

From Learned Space to Founded Space: A MANHAJ Framework for Meaning Measurement in AI

Scientific Research Paper — Paper 4

Author: Faouzi Messeoud — فوزي مسعود

Methodology: MANHAJ — Intellectual Modeling Methodology

Version: 1.0 — October 2026

Status: Scientific Research Paper

About this Paper

This paper applies three laws of the Intellectual Modeling Methodology (MANHAJ) — the Law of Dual Direction of Examination, the Coordinate Origin Law, and the Law of Rightness — to the measurement of meaning in artificial intelligence. It establishes the conditions under which a measurement frame for meaning is founded rather than implicit, formalizes those conditions computationally, and specifies the experimental designs through which their consequences can be tested.

The paper is part of the MANHAJ research program. It continues Paper 02, which established an a priori stage for meaning production in AI and identified the operational representation of meaning states as an open problem, and it draws on Paper 03, which examined the relation between the origin point and the subject in the humanities.

Relationship to the MANHAJ Reference Document

This paper develops a specific research dimension of the methodology presented in: Messeoud, Faouzi. *Intellectual Modeling Methodology (MANHAJ) — Reference Document — Version 1.0*. Zenodo, 15 September 2026. DOI: 10.5281/zenodo.22769627. The present paper is an independent scientific study and should not be considered a replacement for the reference document.

Citation

Messeoud, Faouzi. *From Learned Space to Founded Space: A MANHAJ Framework for Meaning Measurement in AI*. Scientific Research Paper, Paper 4, Version 1.0, 2026.

The scientific paper begins on the following page.

Abstract

Meaning in artificial intelligence is measured through values — similarity scores, distances, drift indices, evaluation ratings — whose reference is rarely examined as a condition of their interpretation. The reference relative to which such a value is read is frequently implicit in the learned representation, or derived from the very outputs being measured. A value obtained under these conditions can be computed with full precision while its conditions of interpretation remain unfounded.

This paper addresses the problem at its source through three laws of the Intellectual Modeling Methodology (MANHAJ). The **Law of Dual Direction of Examination** requires the measurement frame to be founded before measurement, through the order doctrinal centrality → conceptual domain → coordinate origin → coordinates of meaning. The **Coordinate Origin Law** requires measurement to be referenced to an explicit origin independent of the measured property. The **Law of Rightness** requires that an inadmissible frame yield no valid measurements. Together these laws replace measurement whose conditions are implicit with measurement whose conditions are founded, declared, and examined.

The paper translates these requirements into a formal model of a *founded space*: a frame constructed from declared reference and pole formulations, read through a learned encoder, and subjected to a two-level admissibility test that separates methodological admissibility from computational operability. Two formal results follow. Transitions between meaning states are invariant to the choice of origin, whereas deviations from the reference are not; and a reference derived from the measured outputs is structurally blind to the common component of their drift, whatever its magnitude. The paper further establishes, as a methodological consequence of MANHAJ, that every learned representation operates under an embedded doctrinal centrality, and it provides a relational procedure for characterizing the content of that centrality.

Four experimental designs specify how each capability can be tested empirically, each with a strong baseline, independent human ground truth, and an ablation that isolates the contribution of MANHAJ's requirements from that of the implementation technique; their operational protocol is given in an appendix.

Keywords: MANHAJ; artificial intelligence; meaning measurement; doctrinal centrality; conceptual domain; coordinate origin; Law of Rightness; a priori examination; learned representations; evaluation; drift.

1. Introduction

1.1. The problem in AI terms

Contemporary AI systems represent language as vectors in learned spaces, and much of what is said about meaning in these systems is said through quantities computed in those spaces: the similarity of two outputs, the drift of a model across versions, the score an evaluator assigns to a response. These quantities are computed precisely. What is rarely examined is the reference relative to which they are read.

Three recognized problems in current AI practice share this structure.

- **Model-based evaluation.** Language models are increasingly used to judge the outputs of other language models. When the judging reference is derived from systems close to those being judged, the evaluation can favour what it already resembles; documented self-preference effects illustrate the risk (Zheng et al., 2023; Panickssery et al., 2024).

- **Collective drift.** Models trained on generated data can drift collectively, with the tails of the original distribution progressively disappearing (Shumailov et al., 2024). When drift is monitored version-to-version, or relative to the centroid of current outputs, a displacement shared by all outputs can pass unregistered.
- **Compromised references.** Evaluation results are computed and reported on references whose integrity is not established, as in the contamination of benchmarks by training data (Sainz et al., 2023). The metric is computed correctly; its foundation is not examined.

In each case a value is computed precisely relative to a reference that is implicit, derived from the measured material, or compromised. The value is computationally correct and methodologically undetermined.

1.2. The gap left by Paper 02

Paper 02 of the MANHAJ program established an a priori stage for the meaning-producing process in AI, $A_0 = (R, D, Def, C_a, O)$, and monitored meaning states and their transitions ΔM_t against it. It explicitly identified the operational representation of meaning states as a limitation (Paper 02, Section 12.2): ΔM_t was defined formally, but the space in which M_t is measured — its origin, axes, and calibration — was left unspecified. In the language of coordinates, Paper 02 monitored *relative* change without a declared *absolute* frame.

The present paper supplies that frame, and in doing so it founds the measurement on which Paper 02 relies.

1.3. Central question and thesis

Does founding the coordinate system of meaning a priori — on a recognized doctrinal centrality, a conceptual domain, and a declared origin — change the conditions of measuring meaning in AI systems, and which capabilities follow from that change that reference-implicit measurement does not provide?

The thesis of the paper is that it does, and that the change is produced by three MANHAJ laws: the Law of Dual Direction of Examination (Law 2), the Coordinate Origin Law (Law 4), and the Law of Rightness (Law 3). The representational techniques used to implement these laws are established in AI research. The contribution lies in the conditions the laws impose on their use and in the capabilities those conditions produce.

1.4. Contributions

- **C1 — Founded measurement.** A distinction between reference-implicit and founded measurement of meaning, derived from Laws 2 and 4 and the foundational order of MANHAJ.
- **C2 — A formal result on self-derived references.** A proof that a reference derived from the measured outputs is structurally blind to the common component of their drift (Proposition 2).
- **C3 — Foundational invalidity.** A foundational guard derived from Law 3, which introduces into AI measurement the category of a frame whose measurements are invalid at their source, together with a strict separation between methodological admissibility and computational operability.

- **C4 — Embedded centrality.** The establishment, as a methodological consequence of MANHAJ, that every learned representation operates under an embedded doctrinal centrality, and a relational index through which its orientation becomes measurable.
- **C5 — Experimental designs.** Four experimental designs with strong baselines, independent ground truth, and ablations that attribute each capability to MANHAJ’s requirements, with their operational protocol given in Appendix A.

1.5. Scope

The formal model applies to any encoder that maps text to vectors. The experiments are specified for sentence-level encoders and the outputs of language models encoded by a fixed external encoder. The two conceptual domains used — the family and the academic domain of Paper 02’s university-professor scenario — provide continuity with the MANHAJ program.

2. Related Work

2.1. Concept directions in learned spaces

Directions that correspond to conceptual distinctions can be identified in learned spaces from contrasting sets of examples. Early work derived such directions from pairs of word embeddings (Bolukbasi et al., 2016); concept activation vectors identify directions in a network’s activations from user-supplied concept examples (Kim et al., 2018); linear probes test whether information is linearly recoverable from intermediate representations (Alain & Bengio, 2017); and representation-engineering methods read and steer model behaviour along such directions (Zou et al., 2023). The present paper uses pole-based directions of this kind as an implementation tool.

2.2. The implicit geometry of learned spaces

Learned representations are not centred on a semantically meaningful origin. Contextual representations occupy a narrow anisotropic region of the space (Ethayarajh, 2019), and a common corrective practice subtracts the mean of the representations and removes dominant directions (Mu & Viswanath, 2018). Such post-processing establishes an origin derived from the data themselves. It improves many downstream tasks, but the reference it creates is, by construction, a property of the represented material rather than a declared reference.

2.3. Subspace residuals and out-of-distribution detection

Measuring the part of a representation that lies outside a principal subspace is an established technique for detecting inputs that depart from a reference distribution (e.g., Wang et al., 2022). The departure indicator of Section 6 uses this technique relative to a declared, rather than data-extracted, subspace.

2.4. Model-based evaluation, collective drift, and contamination

The use of language models as evaluators, with its documented position and self-preference effects, has been examined systematically (Zheng et al., 2023; Panickssery et al., 2024). Recursive training on generated data has been shown to produce model collapse (Shumailov et al., 2024). Benchmark contamination has been identified as a threat to the validity of reported evaluations (Sainz et al., 2023).

2.5. What this paper adds

These lines of work provide representational tools and document failures of evaluation. Among the works discussed here, none specifies the conditions under which a measurement reference is admissible; none requires that the reference be founded within a recognized framing and be independent of the measured property; and none treats a compromised reference as invalidating all the measurements derived from it. More broadly, we are not aware of work that grounds the measurement of meaning in AI in an ordered foundation of this kind. These conditions are the contribution of MANHAJ's Laws 2, 4, and 3. The tools are used; the conditions are new.

3. Methodological Foundation

3.1. Doctrinal centrality

Doctrinal centrality is the higher framing within which an intellectual or human field is constituted, and within which concepts, values, relations, and interpretations acquire their position (Paper 01, Section 2.1; Reference Document, Section 6.2). Within MANHAJ it is the highest framing of every rational human existence. It is not selected in the manner of an experimental parameter: it is recognized and identified. It is neither a measurement variable nor a coordinate axis. It stands to a field of inquiry as an operating system stands to an application developed within it: the operating system is not one of the application's variables, yet it determines the conditions under which the application is built and runs.

3.2. The foundational order

MANHAJ examines a field through an order of foundation:

Doctrinal Centrality $\rightarrow$ Conceptual Domain $\rightarrow$ Coordinate Origin $\rightarrow$ Coordinates of Meaning

The conceptual domain is constituted within the centrality, $D = D(\Gamma)$; the coordinate origin is established within the domain; and the coordinates of meaning represent meaning states within the domain so framed (Reference Document, Section 6.3.4; Paper 01, Section 2.2). The arrows express an order of foundation and admission, not a mechanical derivation of each element from the one before it.

3.3. Three centralities in the AI setting

Applying this order to artificial intelligence requires three centralities that must not be conflated:

1. Γ_F — **the centrality of the founded frame:** the framing within which the examining field is situated and relative to which meaning is examined. The founded space is built on Γ_F , which is recognized and documented explicitly.
2. Γ_L — **the centrality embedded in the learned encoder:** the framing carried by the human-produced data from which the encoder was formed. The encoder operates within it without having selected it.
3. Γ_C — **the centrality of the represented content:** the framing of the project, text, or discourse whose meaning is represented. As in Paper 03, it is discovered through examination, not imposed by the examiner.

3.4. The embedded centrality of learned systems

Because doctrinal centrality is the higher framing of every rational human existence, the data from which AI systems learn carry the framing within which they were produced. A system that processes such data a posteriori, without examining their foundation, does not operate outside any centrality. It accepts the embedded framing implicitly and operates under its direction. The existence of I_L in every learned representation is therefore a methodological consequence of MANHAJ's foundational order, not a hypothesis awaiting empirical confirmation.

The scope of this claim follows from how MANHAJ defines doctrinal centrality: by its function, not by its source. Any set of conceptions that provides an explanation of existence, of the individual, and of the individual's roles constitutes a doctrinal centrality, whatever its origin and however elementary its form. A learned system does not produce such conceptions by itself: its representations are formed from conceptions borrowed from a prior framing — through the data, and also through the selection of the data, the design of the model, and the definition of its objectives, all of which are human acts. The claim therefore applies to every learned representation that operates a posteriori, whatever the source of its training data, and not only to representations learned from human-written text. It is a claim of the theoretical framework of MANHAJ, not an empirical finding.

What is open to examination is its *content*: which conceptual distinctions the learned representation realizes, and how its realization is oriented relative to declared centralities. Section 6.15 makes this content measurable.

3.5. The three laws

Law of Dual Direction of Examination (Law 2). A priori examination addresses the foundation, references, and approach of a field; a posteriori examination works within a field whose foundation has been accepted, directing itself to phenomena, details, and results (Reference Document, Section 8.2). The two directions are complementary, but the a priori direction is prior.

Law of Rightness (Law 3). The question of permissibility in founding the approach precedes the question of details: *what is not permissible, its details are not valid* (Reference Document, Section 8.3). Rightness is distinct from scientific validity. Research may use correct tools within a field without having examined the rightness of the field's foundation.

Coordinate Origin Law (Law 4). The position of meaning within a conceptual domain is determined relative to a reference coordinate origin; the deviation of a position P_i from the reference position P_0 is $V_i = P_i - P_0$, measured according to the axes and units established by the study (Reference Document, Section 9.1).

3.6. Origin and subject

Paper 01 (Section 2.2) and Paper 03 (Section 3.3) distinguish three forms of relation between the origin and the subject of inquiry: an origin independent of the subject; an origin linked to the subject's position or state; and an origin linked to a property of the subject, which may become the sole criterion by which the subject is judged. Paper 03 (Section 3.4) shows that circularity arises when the conception adopted of a subject becomes the ultimate reference by which that subject is evaluated, without independent examination of its basis. Section 6 gives this circularity an exact computational form.

4. Conditions of Measurement

4.1. The executable and its operating system

Consider an executable file developed in VB.NET and running under Windows. Three layers are involved:

Layer	In software	In AI meaning measurement
The executable	What is run and observed	Outputs and their learned representations
The development environment and source code	VB.NET	The conceptual domain
The operating system	Windows	The doctrinal centrality

Current practice in measuring meaning works on the executable directly. It computes distances, drifts, and scores on outputs and their representations, and it repairs observed failures one behaviour at a time. This resembles reverse engineering: it can reach effective local repairs, but it does not reach the conditions that produced the behaviour. Interpretability research itself often describes its work as reverse engineering, and the analogy is exact in this respect.

MANHAJ requires the executable to be framed first: in which environment it was developed, and under which operating system it runs. Without that framing, an executable can be patched but not treated methodologically at its root.

The analogy applies in two directions. It applies to the content represented, whose centrality Γ_C must be recognized before its meaning is measured. It applies equally to the AI model itself, which is an executable built under the embedded centrality Γ_L of its data (Section 3.4).

4.2. Relative and absolute coordinates

In computer-aided design, a point can be entered relative to the current position ($@20,40$) or in absolute coordinates ($20,40$). Relative input does not depend on the origin of the drawing; absolute input does. The same distinction holds for meaning:

- a **transition** ΔM_t between successive states is a relative quantity, independent of the origin (Proposition 1, Section 6.13);
- a **deviation** D_t from a declared reference is an absolute quantity, which exists only once an origin has been declared.

Paper 02 monitored meaning with relative quantities. Most monitoring of meaning in AI is likewise relative: output against output, version against version, output against the centroid of outputs. It has no declared absolute frame.

4.3. Given and defined coordinate systems

AutoCAD supplies a World Coordinate System (WCS), given by the software, and allows a User Coordinate System (UCS) to be defined for a particular project, with an origin and axes chosen according to the requirements of that project before drawing begins.

The learned space corresponds to the WCS: a coordinate system that was given, not defined for the inquiry at hand. The founded space corresponds to the UCS: a coordinate system defined before measurement, according to the requirements of the inquiry — that is, according to its centrality and conceptual domain. MANHAJ adds what the UCS alone does not contain: the requirement that the definition of the coordinate system be founded in a recognized centrality and examined for admissibility.

4.4. Reference-implicit and founded measurement

We call **reference-implicit** a measurement whose reference is not examined as a condition of its interpretation: it is undeclared, coincides with the learned structure of the system, or is derived from the measured outputs. Fixed reference sets and declared evaluation criteria are widely used, and the critique of this paper does not concern them as such. It concerns measurement practices that do not explicitly examine how the reference was founded, whether it is coherent with the conceptual domain of the measured meaning, and whether it is independent of the measured property.

We call **founded** a measurement whose frame has been established through the foundational order of Section 3.2 and has passed the admissibility test of Section 6.14.

4.5. Necessary difference in conditions, specific difference in results

The conditions of founded and reference-implicit measurement differ necessarily: this follows from the definitions and from Laws 2 and 4. Their results differ in specified ways. Proposition 1 shows that transitions are identical under any origin. Proposition 2 shows that detection capability diverges between a self-derived reference and an independent one. The experimental designs of Section 8 specify how the further capabilities that founded measurement is expected to provide, beyond reference-implicit measurement, can be tested.

5. From Law to Test

5.0. The central argument

A measurement of meaning can be computed with full precision while its conditions of interpretation remain unfounded: nothing in the measurement establishes what the value refers to within the domain, or whether the frame that produced it was admissible. MANHAJ addresses this problem at its source. Its laws establish that measurement is preceded by foundation, and three of them translate this foundational order into requirements that reference-implicit measurement does not satisfy:

- the **Law of Dual Direction of Examination (Law 2)** requires the measurement frame to be founded before measurement;
- the **Coordinate Origin Law (Law 4)** requires measurement to be referenced to an explicit origin independent of the measured property;
- the **Law of Rightness (Law 3)** requires that an inadmissible frame yield no valid measurements.

These requirements constitute the contribution of the paper. The computational techniques used to implement them — concept directions, subspace projection, residual measures, overlap detection — are established implementation tools. MANHAJ supplies the conditions that govern their use, and these conditions produce the capabilities set out below.

The embedded centrality of learned systems (Section 3.4) frames the whole argument: the measurement frames used in current practice are not neutral, because they inherit the framing of the data from which

they were learned. The question is not whether a frame has a centrality, but whether its centrality is recognized, declared, and examined.

Each claim below is stated at one of three levels:

- **[Methodological]** — what follows from MANHAJ's laws, theories, and foundational order;
- **[Computational]** — how the model implements these requirements and what it provides, including properties proved from the model;
- **[Empirical hypothesis]** — an expected result about performance and practical consequence, which the experimental designs of Section 8 are constructed to test. Claims at this level are not presented as established results; they become empirical findings only once the experiments are carried out.

For each law, the section states the law, the requirement it imposes, the mechanism that implements it, the capability that results, and the experimental design that tests it.

5.1. Law of Dual Direction of Examination (Law 2)

Law. A priori examination addresses the foundation, references, and conditions of a field before a posteriori work proceeds within it (Reference Document, Section 8.2).

Requirement. The measurement frame is founded through the full foundational order:

Doctrinal Centrality $\rightarrow$ Conceptual Domain $\rightarrow$ Coordinate Origin $\rightarrow$ Coordinates of Meaning

The centrality I_F is recognized and documented; the domain $D(I_F)$ is constituted within it; reference formulations, poles, and calibration are derived from the operative definitions within $D(I_F)$; and all of this precedes measurement. **[Methodological]**

Current practice builds measurement frames from available examples or data and interprets the resulting values afterwards. Law 2 reverses this order. Independence of the frame from the measured outputs follows from the reversal, but the requirement is stronger than independence: a frame independent of the outputs but drawn, without declaration, from a domain or centrality other than the declared one does not satisfy Law 2. **[Methodological]**

Mechanism. The founded frame F of Section 6.3, each element of which is traceable to the step of the foundational order from which it derives.

Capability.

- The construction of the frame is not shaped by the measured outputs. **[Computational]**
- Each axis carries a declared conceptual meaning within $D(I_F)$, so that every coordinate value is interpretable in the terms of the domain rather than in the terms of the encoder. **[Computational]**
- Foundationally ordered axes are expected to align with the distinctions of the conceptual domain more closely than axes built from convenience examples or extracted from data. **[Empirical hypothesis — tested by E3]**

5.2. Coordinate Origin Law (Law 4)

Law. The position of meaning within a conceptual domain is determined relative to a reference coordinate origin; deviation is $V_i = P_i - P_0$ (Reference Document, Section 9.1).

Requirement. The origin is explicit and independent of the property being measured. An origin tied to the measured property produces circular measurement (Paper 01, Section 2.2; Paper 03, Sections 3.3–3.4). **[Methodological]**

Current monitoring of meaning in AI is predominantly relative. In the terms of Section 4.2, it works with relative coordinates and has no declared absolute frame. Law 4 supplies that frame.

Mechanism. A declared origin p_0 computed from reference formulations; absolute coordinates M_t^F and deviation D_t relative to p_0 ; transitions ΔM_t^F along declared axes (Sections 6.5–6.9).

Capability.

- Absolute referencing adds a dimension of measurement that relative monitoring lacks: the position of each meaning state relative to a declared reference, alongside its change from the preceding state. Transitions are invariant to the origin (Proposition 1); the origin governs deviation. **[Computational]**
- A reference derived from the measured outputs is structurally blind to the common component of their drift, whatever its magnitude (Proposition 2). A reference independent of the outputs does not suffer this blindness: it registers the common component to the extent that the encoder represents the change and that its projection onto the retained axes, $U^T \bar{\delta}$, is non-zero. Proposition 2 concerns representational displacement; that such displacement corresponds to semantic drift is established independently, by human validation in E2. This is the computational form of the circularity identified by Law 4 and Paper 03. **[Computational]**
- Against a fixed external reference, which is also independent of the outputs, the distinctive capability of the founded frame lies in what Law 4 adds beyond independence: an origin declared within a founded domain, and deviation decomposed along declared axes. The founded frame identifies not only that meaning has moved away from the reference, but along which conceptual dimension of the domain it has moved. **[Computational]**. That this capability yields better conceptual attribution than the strong baseline is an **[Empirical hypothesis — tested by E2]**.
- In naturalistic collective drift, the common component is expected to be a substantial share of total drift, in which case the blindness established by Proposition 2 affects real monitoring practice. **[Empirical hypothesis — tested by E2]**

5.3. Law of Rightness (Law 3)

Law. What is not permissible, its details are not valid (Reference Document, Section 8.3).

Requirement. Paper 02 applied the law to transitions between meaning states. This paper applies it at the prior level of the measurement frame itself. If the frame is not admissible, the measurements produced within it — transitions, deviations, closure assessments — are not valid details, whatever their computational precision. **[Methodological]**

This requirement has no counterpart in current practice, where metrics are computed and reported regardless of the condition of the reference that produced them. MANHAJ introduces into the measurement of meaning the category of **foundational invalidity**: a whole family of results is invalidated at its source, without the need to examine each result individually.

Law and guard. The Law of Rightness determines what cannot be treated as valid: measurements produced within an inadmissible frame. The foundational guard is the computational mechanism that

implements the admissibility conditions specified in this paper. The two stand at different levels. If the guard fails to detect a defect, the law is not thereby contradicted; the failure indicates that the admissibility conditions are incompletely specified, or incompletely implemented, for that defect. The experiments therefore test the guard as an implementation of Law 3, not Law 3 itself. **[Methodological]**

Two kinds of failure. A frame can fail methodologically or computationally, and the model treats these failures as distinct:

Check	Conditions	Nature	Established by
Methodological admissibility $\text{Adm}_M(F)$	C (coherence with $D(\Gamma_F)$), E (explicitness), I_{doc} (documented provenance)	Foundational	Documented examination under a written protocol
Computational operability $\text{Op}_{\text{comp}}(F)$	Q (axis discriminability), I_{comp} (no overlap or near-duplication with outputs), Cal (sufficient calibration material)	Representational	Computation

Adm_M	Op_{comp}	State	Consequence
0	—	Foundational invalidity	All measurements within F are invalid; the frame is returned for revision. [Methodological]
1	0	Admissible, not operative	The foundation is sound, but the encoder does not realize the frame or the data do not permit its operation; measurements within F are not interpretable. The failure lies in the representation, not in the foundation. [Methodological]
1	1	Operative	Measurement proceeds.

The second state is itself a result: it locates a limitation of the learned representation by reference to a sound foundation, which reference-implicit measurement, lacking a founded frame, cannot do.

Mechanism. A foundational guard that evaluates $\text{Adm}_M(F)$ before any encoding and $\text{Op}_{\text{comp}}(F)$ after encoding, and blocks or qualifies measurement according to the table above (Section 6.14).

Capability.

- The guard enforces Law 3 computationally: no measurement is reported as founded from a frame that has not passed admissibility. **[Computational]**
- The guard is expected to detect foundational defects — undeclared provenance and undeclared import from another framing — that standard metrics and established auditing procedures do not reveal, and conclusions drawn within such frames are expected to be unreliable against independent ground truth. **[Empirical hypothesis — tested by E1]**

5.4. Embedded centrality

Foundation. Every learned representation operates under an embedded centrality Γ_L (Section 3.4). **[Methodological]**

Mechanism. Two frames for the same concept, founded on two recognized centralities, are realized through the same encoder; the orientation index $\mathcal{O}$ compares how strongly the encoder realizes the distinctions specific to each, relative to the distinctions they share (Section 6.15).

Capability. The content of Γ_L becomes measurable relationally: the orientation of an encoder's realization between declared centralities is defined by the index $\mathcal{O}$. **[Computational]** That this orientation is non-zero for real encoders and varies across encoders formed from different corpora is an **[Empirical hypothesis — tested by E4]**; attributing that variation causally to training data requires the controlled comparison described in Section 8.6.

5.5. Integration

Law 2 determines when and in what order the measurement frame is founded; Law 4 determines how measurement is referenced within it; Law 3 determines what follows when the frame fails. Doctrinal centrality determines the domain within which all three operate, and its embedded counterpart Γ_L becomes an object of measurement. Together they replace measurement whose conditions of interpretation are implicit with measurement whose conditions are founded, declared, and examined.

Claim	Level
The measurement frame must be founded through the full foundational order	Methodological
The origin must be explicit and independent of the measured property	Methodological
An inadmissible frame yields no valid measurements	Methodological
Every learned representation operates under an embedded centrality	Methodological
The founded frame is not shaped by the measured outputs	Computational
Transitions are origin-invariant; the origin governs deviation	Computational
An output-derived reference is blind to the common component of drift	Computational
The founded frame attributes drift to declared conceptual dimensions	Computational
This attribution exceeds that of a fixed external reference	Empirical hypothesis (E2)
Common drift is substantial in naturalistic settings	Empirical hypothesis (E2)
The guard detects foundational defects that standard checks miss	Empirical hypothesis (E1)
Foundationally ordered axes align with domain distinctions	Empirical hypothesis (E3)
The content of the embedded centrality is characterized relationally	Methodological (examination)
The orientation of an encoder between declared centralities is measured by $\mathcal{O}$	Computational
The orientation varies across encoders formed from different corpora	Empirical hypothesis (E4)

6. Formal Model of the Founded Space

6.1. Purpose and governing principles

This section gives the formal construction through which the requirements of Section 5 are implemented. Two principles govern it.

1. **Doctrinal centrality is a condition, not a variable.** It governs which definitions may enter the frame, but it does not appear as a numerical quantity in any equation (Paper 01, Section 2.1).
2. **Methodological admissibility and computational operability are defined on different objects.** $\text{Adm}_M(F)$ is defined on the foundation, the formulations, and their documented provenance, independently of the encoder. $\text{Op}_{\text{comp}}(F)$ is defined on the encoded formulations and the encoded outputs. This structural separation ensures that the methodological validity of a frame is never inferred from, or confused with, the success of its computational implementation.

6.2. The learned space

Let $\mathcal{X}$ denote the set of possible outputs and formulations, and let

$$\varphi: \mathcal{X} \rightarrow \mathbb{R}^n$$

be a learned encoder. A meaning state in the learned space is $m_t = \varphi(o_t)$. The origin of $\mathbb{R}^n$ is the zero vector, whose position relative to meanings is determined by training and is not declared from any conceptual domain: it is an **implicit origin**. Let $O_T = \{o_1, \dots, o_T\}$ denote the measured outputs available at time T .

6.3. The founded frame

The founded frame is

$$F = (\Gamma_F, R, D(\Gamma_F), Def, \mathcal{E}_B, H, P_0, A, \kappa)$$

where:

- Γ_F is the recognized doctrinal centrality of the frame (Section 6.4);
- R is the reference;
- $D(\Gamma_F)$ is the conceptual domain constituted within Γ_F ;
- Def is the set of operative definitions within $D(\Gamma_F)$;
- $\mathcal{E}_B = (E_0, \{E_k^+, E_k^-\}_{k=1}^d, \mathcal{P})$ is the **construction and calibration set**, in which E_0 contains reference formulations of the reference meaning, E_k^+ and E_k^- contain formulations of the positive and negative poles of axis k , and $\mathcal{P} = \{P_1, \dots, P_m\}$ contains paraphrase sets, each holding lexically different formulations of one declared meaning;
- H is the **validation set** of held-out in-domain formulations, written under the same protocol as $\mathcal{E}_B$;
- P_0 is the coordinate origin, with reading p_0 ;
- $A = \{u_1, \dots, u_d\}$ are the representation axes, in an order declared within Def , with $d \ll n$;
- $\kappa = (\kappa_1, \dots, \kappa_d)$ are the calibration coefficients.

Let $\mathcal{E}_{\text{all}} = \mathcal{E}_B \cup H$.

Construction rule. Every formulation in $\mathcal{E}_{\text{all}}$ is written from Def within $D(\Gamma_F)$ and documented before measurement. P_0 , A , κ , and the thresholds ρ and ε are computed from $\mathcal{E}_B$ alone. H is excluded from every construction and calibration step. No element of the frame is computed from O_T .

Validation. Validation on H is a reported diagnostic, not a construction step. If the frame is revised after its results on H have been examined, H becomes construction material, and a new validation set is written under the same protocol; only a set whose results have not informed the frame is described as held out.

6.4. Doctrinal centrality in the model

Position. Doctrinal centrality stands to the frame as an operating system stands to an application developed within it. Γ_F enters no numerical expression of the model; it governs which definitions, and therefore which formulations, may enter the frame.

Formal role. Let $\mathfrak{D}$ be the set of candidate operative definitions. A centrality Γ induces an admissibility relation on $\mathfrak{D}$,

$$\Gamma \vdash d,$$

read: *the definition d acquires its meaning within the framing Γ* . The conceptual domain is the set of concepts whose operative definitions are admitted by Γ_F ; the operative definitions satisfy $Def \subseteq \{d \in \mathfrak{D} : \Gamma_F \vdash d\}$; and every formulation $e \in \mathcal{E}_{\text{all}}$ realizes some $d \in Def$. The foundational order

$$\Gamma_F \rightarrow D(\Gamma_F) \rightarrow P_0 \rightarrow U, K \rightarrow M_t^F$$

expresses an order of foundation and admission.

Recognition. A procedure *Rec* recognizes the centrality within which a frame's domain is constituted: for the founded frame, the centrality of the examining field; in the relational characterization of Section 6.15, the centralities of the source framings of F_A and F_B , each recognized from its own sources. *Rec* produces three kinds of evidence:

1. **Source framing:** the higher references — texts, codified principles, or established conceptions — within which the domain's concepts are constituted;
2. **Derivation trace:** for each $d \in Def$, the trace by which it is admitted under Γ_F ;
3. **Distinguishing elements:** definitions or distinctions admitted under Γ_F that would not be admitted, or would acquire a different meaning, under an alternative framing.

The third kind of evidence distinguishes recognition from documentation: it shows that the frame carries the specific distinctions of its centrality rather than generic content compatible with any framing.

Misrecognition and embedded import. *Misrecognition* occurs when the definitions actually used derive from a framing other than the declared Γ_F . *Embedded import* occurs when definitions or formulations are shaped by the learned system itself — for example, when Def or $\mathcal{E}_{\text{all}}$ is drafted with the assistance of the system under evaluation — so that Γ_L enters the frame that is meant to examine it.

Indeterminacy. If *Rec* does not establish Γ_F , the frame is not founded. Its measurements may be computed, but only with the status of reference-implicit measurement.

Embedded centrality. By Section 3.4, the encoder φ operates under an embedded centrality Γ_L . Γ_L is not directly observable; its content is characterized relationally in Section 6.15.

6.5. The coordinate origin

The reading of the origin is the mean position of the reference formulations:

$$p_0 = \frac{1}{|E_0|} \sum_{e \in E_0} \varphi(e).$$

The origin of the founded space coincides with this **reading** of the reference meaning, not with the reference meaning itself, which is defined methodologically through R and Def . The origin is declared in its definition and learned in its reading.

6.6. The representation axes

For each axis k , the pole centroids and the raw direction are

$$c_k^+ = \frac{1}{|E_k^+|} \sum_{e \in E_k^+} \varphi(e), \quad c_k^- = \frac{1}{|E_k^-|} \sum_{e \in E_k^-} \varphi(e), \quad a_k = c_k^+ - c_k^-.$$

The raw directions are orthonormalized by the Gram–Schmidt procedure in the declared order:

$$\tilde{u}_k = a_k - \sum_{j < k} \langle a_k, u_j \rangle u_j, \quad u_k = \frac{\tilde{u}_k}{\|\tilde{u}_k\|} \quad (\tilde{u}_k \neq 0),$$

collected in $U = [u_1 \cdots u_d] \in \mathbb{R}^{n \times d}$ with $U^\top U = I$. The **orthogonalized-component ratio** is

$$\gamma_k = \frac{\|\tilde{u}_k\|}{\|a_k\|}, \quad 0 \leq \gamma_k \leq 1, \quad a_k \neq 0.$$

If $a_k = 0$, axis k is undefined; if $a_k \neq 0$ and $\tilde{u}_k = 0$, then $\gamma_k = 0$ and u_k is undefined. In both cases the axis fails condition Q (Section 6.14.2). Orthogonality is a property of the representation system and does not assert the independence of the meanings represented (Paper 01, Section 3.4). The pole-based construction is an implementation technique; MANHAJ does not prescribe it. An undefined axis is never used as an operational coordinate: it does not enter U or K , and no quantity measured within the frame $(M_t^F, \Delta M_t^F, D_t, r_t)$ uses it. The value assigned to such an axis by the floor rule of Section 6.15 serves only the diagnostic analysis of the encoder’s realization, and does not make the axis operational.

6.7. Calibration

The calibration coefficient of axis k is

$$\kappa_k = 1/2 \langle a_k, u_k \rangle.$$

Lemma 1. Whenever $\tilde{u}_k \neq 0$, $\langle a_k, u_k \rangle = \|\tilde{u}_k\|$, and therefore $\kappa_k = 1/2 \|\tilde{u}_k\| > 0$.

Proof. $a_k = \tilde{u}_k + \sum_{j < k} \langle a_k, u_j \rangle u_j$, and u_k is orthogonal to every u_j with $j < k$; hence $\langle a_k, u_k \rangle = \langle \tilde{u}_k, u_k \rangle = \|\tilde{u}_k\|$. ■

On its own axis, the two pole centroids are separated by exactly two units, $\langle a_k, u_k \rangle / \kappa_k = 2$. The unit of each axis is defined by the declared poles of the domain, not by the distribution of the measured outputs:

the methodology establishes the axis, and research within the domain establishes its calibration (Paper 01, Section 4.1).

6.8. The meaning state and the division of functions

With $K = \text{diag}(\kappa_1, \dots, \kappa_d)$, the meaning state in the founded space is

$$M_t^F = K^{-1}U^\top(\varphi(o_t) - p_0) \in \mathbb{R}^d.$$

The mean coordinate of the reference formulations is zero,

$$\frac{1}{|E_0|} \sum_{e \in E_0} K^{-1}U^\top(\varphi(e) - p_0) = K^{-1}U^\top(p_0 - p_0) = 0,$$

so that the origin of the founded space coincides with the reading of the reference meaning, operationalizing the Coordinate Origin Law. M_t^F enters the representation of Paper 01 as its meaning component, $P_t = (t, x, M_t^F)$. The transition between successive states is

$$\Delta M_t^F = M_{t+1}^F - M_t^F = K^{-1}U^\top(\varphi(o_{t+1}) - \varphi(o_t)).$$

Element	Governs
Axes U and calibration K	Transitions ΔM_t^F , in declared dimensions and declared units
Origin p_0	Deviation D_t (Section 6.9) and departure indicator r_t (Section 6.10)

6.9. Deviation from the reference

Since the origin of the founded space has coordinate zero, the deviation of state t from the reading of the reference meaning is

$$D_t = M_t^F.$$

$\|D_t\|$ is the distance from the reference reading in calibrated units, and the components of D_t attribute the deviation to the declared axes. This is the operational form of $V_i = P_i - P_0$ in the Coordinate Origin Law; the symbol D_t avoids confusion with the validity variable V_t of Paper 02.

6.10. The departure indicator

For $\varphi(o_t) \neq p_0$,

$$r_t = \frac{\|(I - UU^\top)(\varphi(o_t) - p_0)\|}{\|\varphi(o_t) - p_0\|},$$

and $r_t = 0$ when $\varphi(o_t) = p_0$. In all cases $0 \leq r_t \leq 1$.

r_t indicates departure from the **declared representation** of the domain. A high value may reflect movement toward meanings outside $D(\Gamma_F)$, or dimensions of $D(\Gamma_F)$ not captured by the declared axes. r_t is an indicator of departure, not a measure of domain membership; its relation to membership is examined in E3.

6.11. Calibrated thresholds

Departure threshold. ρ is a declared upper percentile of $\{r(e): e \in E_0 \cup \bigcup_k (E_k^+ \cup E_k^-)\}$, computed from $\mathcal{E}_B$; its behaviour is reported on H .

Change threshold. ε is a declared upper percentile of

$$\{\|K^{-1}U^\top(\varphi(e) - \varphi(e'))\|: e, e' \in P_i, e \neq e', P_i \in \mathcal{P}\},$$

the transition magnitudes between paraphrases of the same declared meaning. The threshold for “no detected meaning change” is thus derived from declared reformulation rather than assumed.

Calibration conditions. ρ requires $|E_0 \cup \bigcup_k (E_k^+ \cup E_k^-)| \geq N_\rho$; ε requires $|P_i| \geq 2$ for every $P_i \in \mathcal{P}$ and $\sum_i |P_i| (|P_i| - 1)/2 \geq N_\varepsilon$. N_ρ , N_ε , and the percentile levels are declared in advance and examined by sensitivity analysis. If a calibration condition fails, the corresponding threshold is not computable and the frame fails condition *Cal*.

6.12. Classification of transitions

Let in_t denote the condition $r_t \leq \rho$. The boundaries are fixed as follows: a transition between two interpretable states is a **detected meaning change** if $\|\Delta M_t^F\| \geq \varepsilon$, and **no detected meaning change** if $\|\Delta M_t^F\| < \varepsilon$. These two classes are not equivalent to the conditions $\Delta M_t = 0$ and $\Delta M_t \neq 0$ of Paper 02, and are not presented as such. $\|\Delta M_t^F\| < \varepsilon$ does not assert that meaning is unchanged; it asserts that no change is detected at the resolution ε , which is calibrated as the variation produced by reformulating one declared meaning (Section 6.11). The model therefore keeps three cases distinct: *no change* ($\Delta M_t^F = 0$), *change below the detection resolution* ($0 < \|\Delta M_t^F\| < \varepsilon$), and *detected change* ($\|\Delta M_t^F\| \geq \varepsilon$). The first two are reported together as “no detected meaning change”, with the value of $\|\Delta M_t^F\|$ recorded, so that the distinction is never lost.

Relation to Paper 02. Paper 02 tests $\Delta M_t = 0$ at the step where it decides whether a transition produced new meaning. In the founded space this step is carried out at the declared resolution ε : the closure detector of Paper 02 receives the transitions classified as “no detected meaning change” between interpretable states, and a closure established on them is a **closure detected at resolution ε** . This changes neither the function of the closure detector — which still requires a persistent absence of new derivational meaning (Paper 02, Law 7) — nor the Law of Rightness: the transition validity V_t of Paper 02 is evaluated on every transition, whatever its magnitude, so that a change below the detection resolution remains subject to the examination of validity. A state is interpretable if $r_t \leq \rho$, the boundary case $r_t = \rho$ being interpretable.

Law 7 defines the end of meaning production as the attainment of a constant: the meaning function reaches C , and its derivative is zero because it is constant. A zero change between two states is a consequence of this condition, not its definition, since a change may vanish momentarily without a constant having been reached. Closure is therefore assessed on the state: over the persistence window, the meaning state remains at a single constant within the resolution ε , and the transitions classified as “no detected meaning change” are a necessary sign of this condition, not a sufficient one.

in_t	in_{t+1}	$\ \Delta M_t^F \ $	Classification
in	in	$\ \Delta M_t^F \ \geq \varepsilon$	Detected meaning change within the declared representation of $D(\Gamma_F)$
in	in	$\ \Delta M_t^F \ < \varepsilon$	No detected meaning change , referred to the closure detector of Paper 02
in	out	—	Departure from the declared representation; not interpreted as a meaning change within $D(\Gamma_F)$
out	in	—	Re-entry ; not interpreted as a meaning change; M_{t+1}^F becomes the next interpretable state
out	out	—	Outside the declared representation; not interpreted within $D(\Gamma_F)$

This classification determines whether meaning has changed relative to the declared representation. It complements the transition validity V_t of Paper 02, which determines whether a change is valid.

6.13. Formal properties

The encoder φ is fixed throughout this subsection.

Proposition 1 (Origin invariance of transitions). If p_0 is replaced by another declared reading p_0' , with U and K unchanged, then

$$M_t^{F'} = M_t^F + K^{-1}U^T(p_0 - p_0'),$$

and hence

$$\Delta M_t^{F'} = \Delta M_t^F.$$

Proof. The added term does not depend on t and cancels under subtraction. ■

Transitions are relative coordinates and are independent of the origin; deviations are absolute coordinates and depend on it. The effect of a declared change of origin is therefore computable and known, consistent with Paper 01 (Section 2.2), which permits a change of origin provided it is declared.

Proposition 2 (Blindness of an output-derived reference to common drift).

Assumptions. (A1) The encoder φ is fixed. (A2) The axes U and the calibration K are fixed, computed from $\mathcal{E}_B$ only. (A3) Two output sequences of equal length T , indexed by the same $t = 1, \dots, T$, satisfy

$$\varphi(o_t') = \varphi(o_t) + \delta_t, \quad t = 1, \dots, T,$$

with mean drift $\bar{\delta} = \frac{1}{T} \sum_{t=1}^T \delta_t$. (A4) The output-derived reference of each sequence is computed from that sequence itself: $\tilde{p}_0 = \frac{1}{T} \sum_{t=1}^T \varphi(o_t)$ and $\tilde{p}_0' = \frac{1}{T} \sum_{t=1}^T \varphi(o_t')$. (A5) The declared reference reading p_0 is computed from formulations contained in neither sequence, and is therefore the same for both.

Statement. Under (A1)–(A5), in the full representation space:

- (i) relative to the output-derived reference,

$$\varphi(o_t') - \tilde{p}_0' = (\varphi(o_t) - \tilde{p}_0) + (\delta_t - \bar{\delta});$$

- (ii) relative to the declared reference,

$$\varphi(o_t') - p_0 = (\varphi(o_t) - p_0) + \delta_t;$$

- (iii) in founded coordinates, the deviations relative to $\tilde{p}_0$ change by $K^{-1}U^\top(\delta_t - \bar{\delta})$, and the deviations relative to p_0 change by $K^{-1}U^\top\delta_t$; averaged over t , the first change is zero and the second is $K^{-1}U^\top\bar{\delta}$.

Proof. By (A3) and (A4), averaging gives $\tilde{p}_0' = \tilde{p}_0 + \bar{\delta}$, which yields (i). By (A5), p_0 is the same for both sequences, which yields (ii). By (A1) and (A2), the map $K^{-1}U^\top$ is the same linear map for both sequences; applying it to (i) and (ii) and averaging over t yields (iii). ■

Corollaries. (a) $\sum_t (\varphi(o_t) - \tilde{p}_0) = 0$: the mean deviation relative to an output-derived reference is zero by construction. (b) **Uniform translation:** if $\delta_t = \delta$ for all t , deviations relative to $\tilde{p}_0$ are unchanged whatever the magnitude of δ , while transitions are also unchanged; only deviations relative to an independent origin register the translation. (c) **Non-uniform drift:** if the δ_t differ, individual deviations relative to $\tilde{p}_0$ do change, by $\delta_t - \bar{\delta}$; what disappears is their average, which is zero for every drift. An output-derived reference therefore registers the dispersion of a drift around its mean but never the common component $\bar{\delta}$ itself. Accordingly, “blind to the common component” in this paper means blind to $\bar{\delta}$, not insensitive to every individual change.

Conditions of registration. Statements (i) and (ii) hold in the full space: there, the declared reference registers the whole common component $\bar{\delta}$, and the output-derived reference registers none of it. In founded coordinates, the declared reference registers the common component if and only if $U^\top\bar{\delta} \neq 0$ (K being invertible). The component of $\bar{\delta}$ orthogonal to the retained axes does not appear in the coordinates; it contributes only to the departure indicator r_t . Conclusions stated in coordinates therefore concern the projected component $U^\top\bar{\delta}$, not the full drift.

Scope. The proposition concerns representational displacement under a fixed encoder. The encoder is external to the generating systems and fixed across conditions, so that outputs of different generating models, model versions, or self-training generations are all encoded by the same φ and the proposition

applies to them directly. It does not cover settings in which the internal representations of a generating model are themselves used as the encoder. That a representational displacement constitutes semantic drift is established independently, through human validation (Section 8.4). The output-derived reference replaces the origin only: $\tilde{p}_0$ is substituted for p_0 while φ , U , and K are held fixed. This is reference condition R1 of Section 8.2.

Proposition 2 is the computational form of what Paper 03 describes as the third relation between origin and subject: when the reference is derived from the property being evaluated, evaluation loses its capacity to detect a change of that property.

6.14. Admissibility of the frame

6.14.1. Methodological admissibility

$$\text{Adm}_M(F) = C \wedge E \wedge I_{\text{doc}}, \quad C, E, I_{\text{doc}} \in \{0,1\},$$

established by documented examination under a written protocol Π :

- **C (Coherence):** Γ_F is recognized through Rec with its three kinds of evidence; $\text{Def} \subseteq \{d: \Gamma_F \vdash d\}$; every formulation in $\mathcal{E}_{\text{all}}$ realizes a definition in Def ; and no reference material is imported from another framing without declaration. $C = 0$ in cases of indeterminacy, misrecognition, or undeclared import.
- **E (Explicitness):** every set in $\mathcal{E}_{\text{all}}$ is documented, and every formulation is traceable to R and Def through the protocol record.
- **I_{doc} (Documented independence):** the provenance of every definition in Def and every formulation in $\mathcal{E}_{\text{all}}$ is recorded, and none was produced by the system under evaluation, derived from its outputs, or drafted with its assistance.

Adm_M is defined on $(\Gamma_F, R, D(\Gamma_F), \text{Def}, \mathcal{E}_{\text{all}})$ and the provenance records. It does not depend on φ or on O_T .

6.14.2. Computational operability

$$\text{Op}_{\text{comp}}(F) = Q \wedge I_{\text{comp}} \wedge \text{Cal}, \quad Q, I_{\text{comp}}, \text{Cal} \in \{0,1\},$$

all computed.

Q (Axis discriminability of the operative frame). For each axis with $\tilde{u}_k \neq 0$ and $|E_k^\pm| \geq 2$, let $\sigma_k^\pm$ be the standard deviations of $\{\langle \varphi(e) - c_k^\pm, u_k \rangle : e \in E_k^\pm\}$, and define the orthogonalized separation ratio

$$S_k^\perp = \frac{\langle a_k, u_k \rangle}{\sigma_k^+ + \sigma_k^-},$$

defined when $\sigma_k^+ + \sigma_k^- > 0$, and $S_k^\perp = +\infty$ when $\sigma_k^+ + \sigma_k^- = 0$.

Then

$$Q = \bigwedge_{k=1}^d (a_k \neq 0 \wedge \tilde{u}_k \neq 0 \wedge \gamma_k \geq \gamma_{\min} \wedge S_k^\perp \geq S_{\min}).$$

The conditions $a_k \neq 0$ and $\tilde{u}_k \neq 0$ exclude degenerate cases; the operative content of Q lies in the thresholds $\gamma_{\min}$ and $S_{\min}$. Every axis falls into exactly one of the following states, and the state is recorded for each axis:

Axis state	Condition	Contribution to Q	Meaning
Discriminating	Defined, sufficient material, $\gamma_k \geq \gamma_{\min}$ and $S_k^\perp \geq S_{\min}$	Satisfies Q for axis k	The axis is operational
Below threshold	Defined, sufficient material, a threshold not met	$Q = 0$	The encoder does not discriminate the axis well enough for operational use
Undefined	$a_k = 0$, or $\tilde{u}_k = 0$	$Q = 0$	No operational direction exists
Insufficient material	Fewer than two formulations in a pole, so that $\sigma_k^\pm$ cannot be computed	$Q = 0$	Discriminability is not determined ; the frame is completed before it can be assessed

$Q = 0$ in the last case expresses that operability has not been established, not that the axis has been shown to be non-discriminating. The case concerns the frame's construction material and is resolved by completing it, not by revising the axis. In every case where $Q = 0$, the relational characterization of Section 6.15 remains available, since it requires $\text{Adm}_M = 1$ and not $\text{Op}_{\text{comp}} = 1$ (Section 6.14.3).

I_{comp} (**Computational independence**). With $\text{dup}(e, o)$, with $0 \leq \text{dup}(e, o) \leq 1$, a declared near-duplication measure (for example, a combination of lexical n-gram overlap and embedding similarity),

$$I_{\text{comp}} = [\mathcal{E}_{\text{all}} \cap \mathcal{O}_T = \emptyset] \wedge \left[\max_{e \in \mathcal{E}_{\text{all}}, o \in \mathcal{O}_T} \text{dup}(e, o) < \tau \right].$$

The threshold τ targets duplication, not the legitimate semantic proximity of in-domain outputs to in-domain formulations.

Cal (**Calibration computability**). $\text{Cal} = 1$ if the calibration conditions of Section 6.11 hold, and 0 otherwise.

Op_{comp} is defined on $\varphi(\mathcal{E}_{\text{all}})$ and $\mathcal{O}_T$. Q and Cal are evaluated once per frame and encoder; I_{comp} is re-evaluated as outputs accumulate. I_{doc} concerns the **origin** of the reference material; I_{comp} concerns **coincidence** between reference material and outputs. A frame may satisfy I_{doc} and fail I_{comp} , for example when the evaluated system reproduces reference text encountered during its training.

6.14.3. States of the frame and the invalidity rule

Adm_M is evaluated first, before any encoding; Op_{comp} after encoding.

$\text{Adm}_M(F)$	$\text{Op}_{\text{comp}}(F)$	State	Status of measurements within F
0	not evaluated	Foundational invalidity	Invalid as founded measurements. If $C = 0$ by indeterminacy, measurements may be computed only with the status of reference-implicit measurement.
1	0	Admissible, not operative	Not interpretable within F . The foundation is sound; the failure lies in the representation (Q), in coincidence with the outputs (I_{comp}), or in insufficient calibration material (Cal).
1	1	Operative	Measurements proceed.

Invalidity rule (Law 3). If $\text{Adm}_M(F) = 0$, every quantity computed within F — M_t^F , ΔM_t^F , D_t , r_t , the transition classification, and any closure assessment — is invalid as a methodological detail, regardless of its computational precision. If I_{comp} fails at time T , measurements involving outputs from that point onward are not interpretable until the coincidence is resolved.

Measurements within a frame and measurements of its realization. Measurements *within* the frame are M_t^F , ΔM_t^F , D_t , r_t , and the transition classification. Measurements *of the frame's realization by the encoder* are the realization profile $\mathcal{R}(\varphi; F)$, the realization classification, and the orientation index $\mathcal{O}$ (Section 6.15). The non-interpretability of measurements in an admissible but non-operative frame concerns measurements within the frame; it does not prevent measuring how the encoder realizes the frame, since failure of Q is precisely what such measurement records. The relational characterization therefore requires $\text{Adm}_M = 1$ and does not require $\text{Op}_{\text{comp}} = 1$. Adm_M , Op_{comp} , and $(\mathcal{R}, \mathcal{O})$ measure three distinct things: methodological admissibility, computational operability, and the relation between frame and encoder.

6.15. Relational characterization of the embedded centrality

Frames and axes. Let Γ_A and Γ_B be two centralities, each recognized through *Rec*, and let F_A and F_B be two frames for the same concept founded on them, each with $\text{Adm}_M = 1$. Neither centrality is ranked above the other or imposed on the other: the characterization describes the encoder relative to the two framings. The axes of each frame are partitioned into

- **shared axes** $\mathcal{A}_{\text{sh}}$: distinctions specified once, by a definition admitted under both centralities with the same pole content, and formulated **separately within each frame** under the same protocol, so that they capture frame-level effects of writing quality and style;
- **distinguishing axes** $\mathcal{A}_{\text{dist}}$: distinctions derived from the distinguishing elements of each centrality, with no counterpart in the other frame.

Data levels. A *distinction* k is identified once. An *axis instance* (k, f) is distinction k within frame f . A *formulation set* (k, f, w) is the set of pole formulations written by writer group w for axis instance (k, f) .

Raw separation. The raw separation is defined on the raw direction of each axis instance, independently of the order of orthogonalization. Three quantities are kept distinct:

- the **raw difference vector** $a_k = c_k^+ - c_k^-$ (Section 6.6), which is *not* normalized; its length $\|a_k\|$ is the distance between the two pole centroids;
- the **unit raw direction** $v_k = a_k / \|a_k\|$, defined when $a_k \neq 0$, with $\|v_k\| = 1$;
- the **pole dispersions** s_k^+ and s_k^- : the standard deviations of the projections $\langle \varphi(e), v_k \rangle$ over $e \in E_k^+$ and over $e \in E_k^-$ respectively, computed from the pooled formulations of both writer groups, with at least the declared minimum per pole.

The numerator of the separation is the distance between the poles along their own direction. Since $\langle c_k^+ - c_k^-, v_k \rangle = \langle a_k, a_k \rangle / \|a_k\| = \|a_k\|$, it can be written equivalently as the difference between the mean projections of the two poles,

$$\|a_k\| = \langle c_k^+, v_k \rangle - \langle c_k^-, v_k \rangle.$$

The raw separation is then:

- if $a_k \neq 0$ and $s_k^+ + s_k^- > 0$:

$$\bar{s}_k = \min\left(\frac{\|a_k\|}{s_k^+ + s_k^-}, S_{\max}\right);$$

- if $a_k \neq 0$ and $s_k^+ + s_k^- = 0$: $\bar{s}_k = S_{\max}$;
- if $a_k = 0$: $\bar{s}_k = S_{\text{floor},k}$.

The numerator uses the unnormalized a_k , never v_k ; the unit direction enters only through the projections that define $s_k^\pm$. The same definition is used for the observed value, for every permuted value in the realization test, and in the orientation index.

$S_{\max}$ is a declared cap, and $S_{\text{floor},k}$ is defined by the floor rule below. Every value satisfies $0 < \bar{s}_k \leq S_{\max}$, so that $\log \bar{s}_k$ is finite.

Realization of an axis. An axis instance is **realized** if the separation of its poles is statistically significant under a permutation test that respects the way the formulations were produced, with control of the false discovery rate across the axis instances of an encoder. The test design is specified in Appendix A.7. v_{sh} and v_{dist} denote the proportions of realized shared and distinguishing axis instances. The realization classification is reported alongside the orientation index and does not enter it.

Three conditions kept distinct. The floor rule separates three conditions that do not imply one another:

- **Zero observed separation:** $a_k = 0$, i.e. the two pole centroids coincide. This does not imply that the encoder merges the formulations within each pole.
- **Collapse in φ :** the encoder maps all formulations of the axis instance to numerically the same point, defined by an independent observable condition,

$$\max_{e, e' \in E_k^+ \cup E_k^-} \|\varphi(e) - \varphi(e')\| \leq \eta_{\text{coll}},$$

where η_{coll} is a declared numerical tolerance. Collapse is established only by this condition, never inferred from the null distribution.

- **Non-estimable floor:** the permutation null distribution of the axis instance does not support an estimate of the chance level — its median is zero, or fewer than N_{null} valid null values remain. This is a property of the test under the protocol, not evidence about the encoder.

Value of an axis with zero separation. An axis instance with $a_k = 0$ receives a declared chance-level value $S_{\text{floor},k}$, estimated from its own permutation null distribution or, failing that, from the pooled null distributions of both frames. The rule, and the cases in which it cannot be applied, are specified in Appendix A.8. The value is diagnostic only: it enters the realization profile and the orientation index, and it does not turn an undefined axis into an operational one (Section 6.6).

Realization profile. The realization profile of the encoder relative to a frame is

$$\mathcal{R}(\varphi; F) = (\{\bar{S}_k\}_k, \{\gamma_k\}_k, \bar{r}_H, \nu_{\text{sh}}, \nu_{\text{dist}}),$$

where $\bar{r}_H$ is the mean departure indicator over H . The components measure different properties: $\bar{S}_k$ the **realization** of each distinction; γ_k the **structure** of the frame; $\bar{r}_H$ the **coverage** of the domain; and $\nu_{\text{sh}}, \nu_{\text{dist}}$ the proportion of distinctions realized at all.

Orientation index. The orientation index aggregates raw separations by geometric means within each cell (frame $\times$ axis type), computed as arithmetic means of $\log \bar{S}_k$ over all axis instances of the cell. Every axis instance, realized or not, enters the comparison with its value $\bar{S}_k$; no axis is excluded, so that no frame appears better realized through the removal of distinctions the encoder represents weakly. The distinguishing realization score of a frame is

$$\text{Real}(\varphi; F) = \frac{1}{|\mathcal{A}_{\text{dist}}|} \sum_{k \in \mathcal{A}_{\text{dist}}} \log \bar{S}_k - \frac{1}{|\mathcal{A}_{\text{sh}}|} \sum_{k \in \mathcal{A}_{\text{sh}}} \log \bar{S}_k,$$

and the orientation index is

$$\mathcal{O}(\varphi; F_A, F_B) = \text{Real}(\varphi; F_A) - \text{Real}(\varphi; F_B) = \log \frac{G_{\text{dist}}^A / G_{\text{sh}}^A}{G_{\text{dist}}^B / G_{\text{sh}}^B},$$

where G denotes the geometric mean of $\bar{S}_k$ over the indicated axes. $\mathcal{O}$ is the interaction between frame and axis type in a 2×2 design. The realization classification, and therefore the multiplicity procedure, determines ν_{sh} and ν_{dist} only; it does not enter Real or $\mathcal{O}$.

Conditions of definition. The geometric mean of a cell (frame $\times$ axis type) is defined when the cell contains at least one axis instance and every axis instance in it has a value $\bar{S}_k$, which holds for every instance with $a_k \neq 0$ and for an instance with $a_k = 0$ once its floor has been assigned. $\mathcal{O}(\varphi; F_A, F_B)$ is defined if and only if all four cells — $\mathcal{A}_{\text{dist}}$ and $\mathcal{A}_{\text{sh}}$ in F_A and in F_B — are defined. It is used as a primary result only if, in addition, each cell contains at least the declared minimum number n_{cell} of axis instances, after the exclusion of shared pairs that fail the equivalence judgment (Appendix A.6). The index is therefore in exactly one of four states: **defined and primary**; **defined but descriptive** (a cell below n_{cell} , or pooled-floor values above the declared ceiling of Appendix A.8); **not computable under the floor rule** (a required floor cannot be assigned); or **undefined** (an empty cell). The state is always reported with the index, and only the first supports a characterization of Γ_L .

Weak distinguishing axes. Because every axis enters with its value, weakly realized or unrealized distinguishing axes lower $\text{Real}(\varphi; F)$ by their actual magnitude, which lies near the chance level $S_{\text{floor},k}$. $\mathcal{O}$ therefore remains interpretable in this case: a negative contribution from a frame's distinguishing axes records that the encoder flattens that frame's specific distinctions. A value near zero, $\mathcal{O} \approx 0$, alone does not determine its interpretation, because it can arise in more than one way: the distinguishing axes of both frames may be weak while shared axes are strong (flattening of both framings), or the separations of the two frames may simply be comparable at any level (no differential orientation). The two cases are distinguished only by the separation levels in each of the four cells — the geometric means $G_{\text{dist}}^A, G_{\text{sh}}^A$,

$G_{\text{dist}}^B, G_{\text{sh}}^B$ — together with v_{sh} and v_{dist} . These cell-level values are always reported with $\mathcal{O}$, and a value of $\mathcal{O}$ is never interpreted without them.

Levels of the claim. Four claims about embedded centrality are kept distinct throughout the paper: (1) the *existence* of Γ_L in every learned representation is a methodological consequence of MANHAJ (Section 3.4); (2) its *content* is a matter of examination and characterization; (3) the *orientation* of an encoder's realization between declared centralities, measured by $\mathcal{O}$, is a computational result defined by the model; (4) the attribution of that orientation to the encoder's *training data* is a causal conclusion, which variation across encoders does not establish on its own, since encoders also differ in architecture, dimensionality, objectives, and preprocessing; it requires a controlled comparison of encoders that are close in architecture and objectives but differ in documented training data.

Interpretation. $\mathcal{O} > 0$ establishes a relative orientation of the encoder's realization toward the distinctions specific to F_A , compared with those specific to F_B , for the concepts and axes examined. It characterizes Γ_L relative to a declared pair of centralities; it does not identify Γ_L in full. A low realization of the distinctions of a frame bears on the encoder, not on the admissibility of the centrality on which the frame is founded, which is established by Adm_M independently of φ . An orientation of zero does not bear on the existence of Γ_L , which is established methodologically.

6.16. Unrealized frames as a diagnostic state

The state ($\text{Adm}_M = 1, \text{Op}_{\text{comp}} = 0$) with failure of Q identifies distinctions required by a methodologically sound domain that the encoder does not realize. Because the foundation is established independently of φ , the failure is attributable to the representation rather than to the frame. The founded frame thus functions as a diagnostic instrument for the learned representation, and contributes to the relational characterization of Γ_L .

6.17. A priori definition and a posteriori readings

The founded frame is a priori in its definition — the centrality, domain, reference, poles, and units of measurement — and a posteriori in its readings, since $p_0, c_k^\pm, U$, and $\varphi(o_t)$ are computed through the learned encoder. The foundation of the learned space is implicit; the foundation of the founded space is recognized, declared, and examined, with its methodological admissibility established independently of its encoding. The learned encoder supplies the raw material of measurement; MANHAJ supplies the frame.

6.18. Sequence of the model

Step	Operation	Failure outcome
1	Recognition of Γ_F through <i>Rec</i>	Indeterminacy: reference-implicit status
2	$D(\Gamma_F), R, Def \subseteq \{d: \Gamma_F \vdash d\}$	—
3	Declared formulation sets $\mathcal{E}_B$ and H , with provenance records	—
4	$\text{Adm}_M(F) = C \wedge E \wedge I_{\text{doc}}$	Foundational invalidity
5	Encoding of $\mathcal{E}_B$: $p_0, a_k, \tilde{u}_k, u_k, \gamma_k, \kappa_k, \rho, \varepsilon$	—
6	$\text{Op}_{\text{comp}}(F) = Q \wedge I_{\text{comp}} \wedge Cal$	Admissible, not operative (realization still measured)
7	Validation on H (reported diagnostic)	—
8	$M_t^F, D_t, r_t, \Delta M_t^F$, transition classification	—
9	Closure detection and transition validity (Paper 02)	—

6.19. Correspondence between MANHAJ and the formal elements

Methodological source	Formal elements
Theory of Doctrinal Centrality	$\Gamma \vdash d$; $D(\Gamma_F)$; <i>Rec</i> ; condition C ; embedded centrality Γ_L ; $\mathcal{R}(\varphi; F)$ and $\mathcal{O}$
Law of Dual Direction of Examination (Law 2)	Construction rule (Section 6.3); evaluation of Adm_M before any encoding
Coordinate Origin Law (Law 4)	p_0 ; M_t^F with origin at zero; D_t ; I_{doc} and I_{comp} ; Propositions 1 and 2
Law of Rightness (Law 3)	States of the frame and invalidity rule (Section 6.14.3)
Theory of Meaning (Reference Document, Section 6.3.4)	M_t^F ; r_t ; departure and re-entry in the transition classification

6.20. Notation

Symbol	Meaning
$\Gamma_F, \Gamma_L, \Gamma_C$	Centralities of the founded frame, the learned encoder, and the represented content
Γ_A, Γ_B	Two declared centralities in the relational characterization
$\Gamma \vdash d$	Definition d is admitted within the framing Γ
<i>Rec</i> , Π	Recognition procedure; written protocol for methodological examination
φ, n, O_T	Learned encoder; dimension of its space; measured outputs at time T

Symbol	Meaning
$\mathcal{E}_B, H, \mathcal{E}_{\text{all}}$	Construction-and-calibration set; validation set; their union
$E_0, E_k^\pm, \mathcal{P}$	Reference, pole, and paraphrase formulation sets
p_0	Reading of the reference meaning (origin of the founded space)
a_k, v_k	Raw (unnormalized) difference vector of the poles; its unit direction
$\tilde{u}_k, u_k, U$	Orthogonalized and normalized axis directions; axis matrix
$s_k^\pm$	Pole dispersions along v_k (raw separation)
d, γ_k	Number of axes; orthogonalized-component ratio
κ_k, K	Calibration coefficients; calibration matrix
$M_t^F, \Delta M_t^F, D_t$	Meaning state, transition, and deviation in the founded space
r_t, ρ, ε	Departure indicator; departure threshold; change threshold
N_ρ, N_ε	Minimum calibration material for ρ and ε
$S_k^\perp, S_{\min}, \gamma_{\min}$	Orthogonalized separation ratio and thresholds of condition Q
$\bar{S}_k, S_{\max}, S_{\text{floor},k}$	Raw separation; its cap; its floor
$N_{\text{null}}, N_{\text{pool}}, \lambda_{\text{pool}}, \eta_{\text{coll}}$	Minimum valid null values; minimum axis-specific medians for the pooled floor; ceiling on pooled-floor values per cell; collapse tolerance
B, m, m_w, q, H_m	Permutations; axis instances in the test family; writers per axis instance; FDR level; harmonic sum
$\mathcal{A}_{\text{sh}}, \mathcal{A}_{\text{dist}}, n_{\text{cell}}$	Shared and distinguishing axes; minimum axis instances per cell
$\mathcal{R}(\varphi; F), \bar{r}_H, v_{\text{sh}}, v_{\text{dist}}$	Realization profile; mean departure on H ; proportions of realized axes
$\text{Real}(\varphi; F), \mathcal{O}(\varphi; F_A, F_B)$	Distinguishing realization score; orientation index
dup, τ	Near-duplication measure; its threshold
$\text{Adm}_M, C, E, I_{\text{doc}}$	Methodological admissibility and its conditions
$\text{Op}_{\text{comp}}, Q, I_{\text{comp}}, \text{Cal}$	Computational operability and its conditions
V_t	Transition validity (Paper 02)

7. Integration with Paper 02

The founded space completes the computational framework of Paper 02 at four points.

1. **Foundation before process.** The guard of Section 6.14 precedes the a priori state $A_0 = (R, D, Def, C_a, O)$ of Paper 02: before the validity of meaning production is examined, the

admissibility of the frame in which meaning is measured is established. Law 3 thus operates at two levels — the frame (this paper) and the transition (Paper 02).

2. **Relative monitoring with declared dimensions.** ΔM_t of Paper 02 becomes ΔM_t^F , expressed along declared axes, in declared units, with a change threshold ε calibrated from declared paraphrase.
3. **Absolute reference.** The deviation D_t from the declared reference, absent from Paper 02, is added. Paper 02's monitoring thereby gains the absolute coordinates that relative monitoring lacks.
4. **Computable closure.** Closure in the sense of Law 7 — the attainment of a cognitive constant C — becomes computable within the founded space at the declared resolution ε : the meaning state is shown to remain at a constant over the persistence window, the transitions classified as “no detected meaning change” in Section 6.12 being the necessary sign of this condition. This is a closure *detected at resolution ε* , not an assertion that $\Delta M_t = 0$ exactly; departures and re-entries are separated from it, so that a change of frame is not mistaken for closure or for new meaning. Law 7 thereby gives the derivational processing of an AI system a founded stopping criterion: processing ends when it no longer produces new derivational meaning within the declared domain, rather than continuing indefinitely or stopping at an arbitrary limit.

The resulting sequence extends that of Paper 02:

Centrality $\rightarrow$ Founded Frame $\rightarrow$ Admissibility $\rightarrow$ Meaning State $\rightarrow$ Transition $\rightarrow$ Validity $\rightarrow$
Closure

8. Experimental Designs

8.1. Status and principles

This paper is a theoretical and methodological contribution. Its claims rest on MANHAJ's laws and on the formal model; the experiments below are **designs** that specify how each capability derived from the laws can be demonstrated empirically. Their operational protocol — annotation, sampling, test procedures, statistical analysis, and declared constants — is set out in Appendix A. Proof-of-concept implementations are part of subsequent work in the MANHAJ program.

Four principles govern the designs. (1) **Strong baseline:** the founded frame is compared not only with output-derived and relative references but with a fixed external reference, so that what the experiments test is the contribution of MANHAJ's requirements, not of reference fixity alone. (2) **Independent ground truth:** semantic changes and domain distinctions are established by human judgment independent of the measures under test. (3) **Pre-specified expectations:** the expected result of each experiment is stated in advance. (4) **Attribution through ablation:** a condition that shares the implementation technique but not the foundational order isolates the contribution of the laws.

8.2. Reference conditions

Code	Reference condition	Independent of outputs	Founded under MANHAJ
R1	Output-derived centroid, substituted for p_0 with φ , U , and K held fixed	No	No
R2	Version-to-version (relative) comparison	No	No
R3	Strong baseline: fixed external set of curated, human-written in-domain texts, compared by standard similarity	Yes	No
R3a	Ablation: fixed external reference with axes built by the same technique as R4 from convenience examples	Yes	No
R4	Founded frame	Yes	Yes

R3a shares R4’s implementation technique and differs from it only in the foundational order. The comparison of R4 with R3a therefore isolates the contribution of MANHAJ’s requirements.

8.3. Experiment E1 — The foundational guard (Law 3)

Question. Does applying the Law of Rightness at the level of the measurement frame detect defects that conventional checks leave unrevealed?

Design. Sound frames are compared with frames carrying one known defect each: contamination of the reference by the outputs, reference material produced by the evaluated system, undeclared import from another framing, and non-discriminative axes. Standard metrics, established auditing procedures, and the foundational guard are applied to every frame, and the conclusions drawn in each frame are compared with ground truth (Appendix A.3).

Expected result. Representational defects (contamination, non-discriminative axes) are detected by the guard’s computational checks and may also be detected by established auditing procedures. **Foundational defects** (provenance, undeclared import) leave standard metrics and conventional checks unaffected, because they concern the origin and coherence of the reference rather than its surface properties; they are detected by the methodological check Adm_M . Conclusions drawn within frames carrying foundational defects are unreliable against ground truth, although their metrics appear normal.

What it would establish. If the expected result is obtained: that the category of foundational invalidity, introduced by Law 3, provides a safeguard that conventional measurement lacks.

8.4. Experiment E2 — Absolute referencing and drift (Law 4)

Question. Does absolute referencing within a founded frame provide drift detection and conceptual attribution beyond relative and fixed external references, and is the common component of drift significant in practice?

Design. Constructed semantic drift validated by human judgment, meaning-preserving controls (paraphrase, style, length), and naturalistic collective drift across self-training generations or model versions are measured under the reference conditions R1–R4 (Appendix A.4).

Expected result. R1 does not register the common component of drift, as Proposition 2 requires; R2 registers step-to-step change but no cumulative displacement relative to a reference. R4 discriminates semantic drift from meaning-preserving change at least as well as R3, and in addition attributes drift to declared conceptual dimensions — a capability R3 does not possess. R4 exceeds R3a in attribution, so that the advantage derives from the foundational order. The common component is a substantial share of naturalistic drift.

What it would establish. If the expected result is obtained: that the Coordinate Origin Law adds absolute, conceptually decomposed referencing to a practice that works with relative quantities, and that output-derived references fail on a practically significant component of real drift.

8.5. Experiment E3 — The foundational order (Law 2)

Question. Do axes founded through the full foundational order measure the distinctions of the conceptual domain more closely than axes constructed without it?

Design. Founded axes (R4) are compared with axes built by the same technique from convenience examples (R3a) and with axes extracted from data without supervision (R4u), against judgments of independent domain experts on held-out items, with controls that separate the foundational order from expert curation (Appendix A.5).

Expected result. R4 exceeds R3a and R4u in alignment with expert judgments and in separating in-domain from out-of-domain items.

What it would establish. If the expected result is obtained: that the a priori order required by the Law of Dual Direction of Examination produces axes that measure the domain's own distinctions.

8.6. Experiment E4 — Relational characterization of the embedded centrality

Question. How is a learned encoder's realization oriented between two frames founded on different declared centralities, and how does that orientation vary across encoders formed from different corpora?

Design. For two concepts (candidates: *family* and *progress*), two frames are founded on two recognized centralities, each with shared and distinguishing axes. Three encoders formed from different corpora realize both frames, and the orientation index $\mathcal{O}$ of Section 6.15 is computed with its cell-level separations and realization proportions (Appendix A.6).

Expected result. A systematic non-zero orientation on the distinguishing axes while shared axes are realized comparably; stability across writer groups and concepts; and variation of the orientation across encoders, described as covariation with their training corpora. A causal attribution of the variation to training data is not expected from this design, since the encoders also differ in architecture, dimensionality, objectives, and preprocessing. Where encoders that are close in architecture and objectives but differ in documented training data are available, their comparison provides the more controlled design that such an attribution requires.

What it would establish. If the expected result is obtained: that the content of the embedded centrality Γ_L , whose existence MANHAJ establishes methodologically, becomes measurable relationally.

8.7. From results to claims

Result	Contribution the result would establish
E1 as expected	Law 3 applied at the frame level detects foundational defects — provenance and undeclared import — that standard metrics and established auditing procedures do not reveal
E2 as expected	Absolute referencing within a founded frame provides conceptual drift attribution beyond fixed external references, and output-derived references fail on a significant component of real drift
E3 as expected	The foundational order of Law 2 produces axes that measure the domain’s own distinctions
E4 as expected	The orientation of a learned encoder’s realization between declared centralities is measurable, and it varies across encoders formed from different corpora; a causal attribution to training data requires the controlled comparison of Section 8.6
Results differing from expectations	Analysed at the level at which they arise — implementation, tested capability, or scope — and used to refine the operational specification; the methodological requirements of the three laws stand at the methodological level, where they are established

Frames that are methodologically admissible but not operative are analysed individually as a distinct result: they identify distinctions required by a founded domain that the encoder does not realize (Section 6.16).

9. Discussion

9.1. What changes when the frame is founded

The object of methodological examination changes. Reference-implicit measurement examines encoded representations and the relations computed within them. Founded measurement also examines the conditions under which those relations count as measurements of meaning: the centrality within which the domain is constituted, the domain itself, the origin, the axes, and their calibration. The encoder remains indispensable, but the status of its outputs changes. They are read within a declared frame whose admissibility is not determined by the encoder’s numerical success.

9.2. The origin and absolute referencing

Relative quantities — transitions, version-to-version differences, distances to the centroid of current outputs — are the dominant instruments of meaning monitoring in AI. Proposition 1 shows that such quantities are blind to the choice of origin, and Proposition 2 shows that a reference built from the outputs themselves is blind to their common drift. The Coordinate Origin Law supplies what these instruments lack: an absolute position relative to a declared reference, decomposed along declared conceptual dimensions. In the terms of Section 4.2, MANHAJ adds the absolute coordinate to a practice that works with relative coordinates only.

9.3. Foundational invalidity as a safeguard

The Law of Rightness introduces into AI measurement a category that current practice does not possess: a family of results invalid at its source. A contaminated benchmark, a reference written by the evaluated system, or a reference imported from another framing without declaration can produce metrics that are computed correctly and give no sign of the defect; E1 is designed to test the extent to which this is so. The foundational guard does not examine these metrics one by one; it examines the frame that produced them, and its verdict applies to all of them. This is the practical form of the principle that what is not permissible, its details are not valid.

9.4. The executable and its operating system

Current practice treats the outputs of AI systems, and the systems themselves, as executables to be analysed and repaired. MANHAJ requires both to be framed: the content by its centrality Γ_C , and the model by the centrality Γ_L embedded in its data. The relational characterization of Section 6.15 makes the second requirement operational. It turns the question of which framing a learned representation realizes from a matter of assertion into a matter of measurement, while keeping the existence of that framing where MANHAJ places it: at the methodological level.

9.5. Relation to learned representations

The founded space does not replace learned representations. It uses them as the raw material of measurement and places over them a frame they do not supply. This is a division of roles: the learned encoder generates numerical structure; the founded frame establishes the reference conditions under which that structure is read. A learned encoder can therefore be evaluated against a founded frame, and such an evaluation is designed to show where it realizes the declared distinctions and where it does not — information that is unavailable when the representation serves as its own reference.

9.6. Three levels of claims

The paper makes claims at three levels. Methodological claims follow from MANHAJ's laws: the frame must be founded, the origin must be independent, an inadmissible frame yields no valid details, and every learned representation operates under an embedded centrality. Computational claims follow from the model: Propositions 1 and 2, Lemma 1, and the properties of the admissibility functions. Empirical claims are, at this stage, hypotheses: Experiments E1–E4 are designed to test them, and their results are not presented here. Keeping the levels distinct gives each its proper force: the methodological claims and the mathematical results do not wait on experiments, and the experiments, once carried out, will show concretely where and how much the methodological difference matters.

10. Epistemic Status and Scope

- **Tools and conditions.** The representational tools used — concept directions, projection, residuals, overlap detection — are established. The contribution lies in the conditions MANHAJ imposes on their use and in the capabilities those conditions produce.
- **A priori definition, a posteriori readings.** The founded frame is a priori in its definition and a posteriori in its readings, which pass through the learned encoder. The difference between the two spaces is that the foundation of one is implicit while the foundation of the other is declared and examined.

- **Methodological examination.** Conditions C , E , and I_{doc} are established by documented examination under the protocols Rec and Π , which are written as part of the operational protocol before the experiments, with the family domain as their worked example.
- **Formal scope.** Proposition 2 concerns representational displacement under a fixed encoder; only the projected component $U^T \bar{\delta}$ of the common drift appears in the coordinates, and its semantic interpretation rests on human validation.
- **Operational protocol.** The test procedures, thresholds, and sample sizes of Appendix A are implementation decisions; they do not bear on the methodological claims, and proof-of-concept implementations follow in subsequent work.
- **Operational constants.** S_{max} , N_{null} , N_{pool} , λ_{pool} , η_{coll} , q , γ_{min} , S_{min} , τ , N_{ρ} , N_{ϵ} , and the percentile levels are implementation choices, fixed before the experiments and examined by sensitivity analysis.
- **Axis order.** Gram–Schmidt orthogonalization makes the operational axes depend on their declared order; the order is declared within Def , and the relational characterization uses raw directions, which do not depend on it.

11. Conclusion

This paper has examined the foundation of meaning measurement in artificial intelligence. Values that express meaning in AI systems — similarities, drifts, scores — are computed precisely relative to references that are implicit, derived from the measured outputs, or compromised. MANHAJ addresses this problem at its source.

Three laws produce the change. The Law of Dual Direction of Examination requires the measurement frame to be founded before measurement, through the order doctrinal centrality $\rightarrow$ conceptual domain $\rightarrow$ coordinate origin $\rightarrow$ coordinates of meaning. The Coordinate Origin Law requires an explicit origin independent of the measured property, and thereby adds absolute coordinates to a practice that works with relative ones. The Law of Rightness requires that an inadmissible frame yield no valid measurements, and thereby introduces the category of foundational invalidity.

The formal model implements these requirements in a founded space read through a learned encoder. It proves that transitions are invariant to the origin while deviations are not, and that a reference derived from the measured outputs is blind to the common component of their drift. It separates methodological admissibility from computational operability, so that the validity of a frame is never inferred from the success of its implementation. It establishes that every learned representation operates under an embedded centrality, and it makes the orientation of that centrality measurable relative to declared framings. It also makes the derivational closure of Law 7 — the attainment of a cognitive constant — computable at a declared resolution, giving derivational processing a founded stopping criterion.

Four experimental designs, each with a strong baseline, independent ground truth, and an ablation that isolates MANHAJ’s requirements, specify how the resulting capabilities are to be tested: detection of foundational defects that conventional checks leave unrevealed, conceptual attribution of drift, axes that measure the domain’s own distinctions, and a relational characterization of the framing embedded in learned representations.

The distinction between learned space and founded space is therefore not a distinction between two arrangements of vectors. It is a distinction between a reference that is assumed and a reference that is

founded. A representation provides coordinates; a computation provides a result; a founded frame establishes the conditions under which that result is a measurement of meaning.

12. Note on Development and Commercial Use

MANHAJ is an original methodology whose intellectual rights and authorship remain vested in its author.

Any use, integration, adaptation, exploitation, commercialization, or commercial implementation of MANHAJ requires the prior written authorization of the author through a paid commercial license.

No free commercial license is granted. Any authorization to commercially exploit MANHAJ is necessarily subject to the payment of licensing fees and to the contractual conditions established by the author.

The terms of the commercial license, including its scope, duration, field of use, rights granted, financial conditions, attribution requirements, and provisions relating to intellectual property, shall be defined in a contract concluded with the author.

The publication of this scientific paper does not, under any circumstances, constitute a commercial license and does not grant any right to commercially exploit MANHAJ.

No commercial use, commercialization, integration into a product or service, or exploitation of MANHAJ for commercial purposes may be considered authorized solely by virtue of the publication of this paper.

13. References

MANHAJ documents

1. Messeoud, Faouzi. *Intellectual Modeling Methodology (MANHAJ) — Reference Document — Version 1.0*. Zenodo, 15 September 2026. DOI: 10.5281/zenodo.22769627.
2. Messeoud, Faouzi. *MANHAJ: A Generative Methodological Framework for Intellectual Modeling*. Scientific Research Paper, Paper 1, Version 1.0, October 2026. DOI: 10.5281/zenodo.23138923.
3. Messeoud, Faouzi. *Examining Meaning Production in AI: a MANHAJ Framework*. Scientific Research Paper, Paper 2, Version 1.0, October 2026. DOI: 10.5281/zenodo.23214262.
4. Messeoud, Faouzi. *The Epistemological Foundations of Modern Humanities in Light of the Intellectual Modeling Methodology (MANHAJ): Posterior Inquiry, the Origin Point, and the Problem of Generalizing the Historical Model*. Scientific Research Paper, Paper 3, Version 1.0, October 2026. DOI: 10.5281/zenodo.23558985.

Author's earlier works

1. Messeoud, Faouzi. *The Immaterial Link of Tunisians with France: How France Sustains Tunisian Dependency*. First edition. Dar El Alaoui, 2023. ISBN 978-9938-9684-5-3.
2. Messeoud, Faouzi. *Doctrinal Centrality and the Conceptual Field: A Manifesto on Methodological Corruption and the Foundations of Subjugation*. Dar Wachma (Tunisia), 2024. ISBN 978-9938-9740-6-5.

Artificial intelligence and representation

- Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *ICLR 2017 Workshop Track*. arXiv:1610.01644.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems* 29.
- Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877–1901.
- Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of EMNLP-IJCNLP 2019*, 55–65.
- Golchin, S., & Surdeanu, M. (2024). Time travel in LLMs: Tracing data contamination in large language models. *International Conference on Learning Representations (ICLR 2024)*. arXiv:2308.08493.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, 2668–2677.
- Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. *International Conference on Learning Representations (ICLR 2018)*. arXiv:1702.01417.
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37. arXiv:2404.13076.
- Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. *Findings of the Association for Computational Linguistics: EMNLP 2023*. arXiv:2310.18018.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Wang, H., Li, Z., Feng, L., & Zhang, W. (2022). ViM: Out-of-distribution with virtual-logit matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 4921–4930.
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*.
- Zou, A., Phan, L., Chen, S., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.

Statistics

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), 1165–1188. <https://doi.org/10.1214/aos/1013699998>

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). SAGE Publications.

Owen, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2), 386–411. <https://doi.org/10.1214/07-AOAS122>

Appendix A. Operational Protocol (proposed)

A.1. Status of this appendix

This appendix specifies the operational protocol through which the experimental designs of Section 8 can be carried out. It belongs to the implementation level: it does not add to the methodological claims of the paper, and its choices — test procedures, thresholds, sample sizes — are implementation decisions fixed before any experiment and examined by sensitivity analysis. The values of the declared constants, and the recognition and examination protocols Rec and Π , are fixed in the protocol document that precedes the experiments.

A.2. Setup and annotation

- **Encoder.** One primary sentence-level encoder, external to the generating systems and fixed across conditions, with a second encoder for robustness. Experiment E4 uses three encoders (Appendix A.6).
- **Generating systems.** One language model producing outputs; self-training generations or successive model versions as specified in E2, all encoded by the same fixed encoder.
- **Domains.** The family domain of Paper 02, which provides a concrete case for documenting condition C , and the academic domain of Paper 02's university-professor scenario.
- **Frame documentation.** For each domain: the recognition record of Γ_F produced by Rec ; the determination of $D(\Gamma_F)$; the operative definitions Def ; the sets E_0 , $E_k^\pm$, $\mathcal{P}$, and H ; and a protocol record linking each element to its step in the foundational order.

Annotators are blind to reference conditions and measured values. Their tasks are: presence and magnitude of semantic change; direction of change relative to declared domain distinctions; meaning preservation in control items; and domain membership of held-out items. Agreement is reported with Krippendorff's α (Krippendorff, 2018); items below a pre-specified threshold are analysed separately.

A.3. Experiment E1 — procedure

Frames. Sound frames, founded under the full procedure; and compromised frames, each carrying one known defect:

- (a) **contamination:** reference formulations overlapping or near-duplicating measured outputs (I_{comp});
- (b) **provenance:** reference formulations produced by the evaluated system (I_{doc});
- (c) **undeclared import:** reference material from another domain or centrality without declaration (C);
- (d) **non-discriminative axes** (Q).

Conventional checks. The following established procedures are applied to every frame, sound and compromised, and their results are reported per defect type:

- **n-gram overlap** between reference formulations and measured outputs (13-gram overlap, as used in contamination analyses of large language models; Brown et al., 2020);
- **embedding near-duplicate detection** between reference formulations and outputs, with a declared similarity threshold;

- **guided-instruction contamination detection**, which tests whether the evaluated system completes reference formulations it has encountered (Golchin & Surdeanu, 2024);
- **distributional checks** of the standard metrics themselves: comparison of their distributions between sound and compromised frames.

Procedure. (1) Compute standard metrics — similarity, drift, change detection — within each frame. (2) Apply the conventional checks above. (3) Apply the guard: $\text{Op}_{\text{comp}}(F)$ computationally; $\text{Adm}_M(F)$ by two independent reviewers applying the written protocol, blind to the frame’s construction.

Information available to each procedure. The comparison tests the procedures, not differences in the information given to them. Each procedure receives the same frame materials: the formulation sets, the measured outputs, and the frame documentation — including the provenance records and the recognition record of *Rec* — when the procedure is able to use them. The defects are of two kinds, analysed separately: **recorded defects**, whose trace is present in the documentation (for example, a provenance record showing that formulations were produced by the evaluated system), and **unrecorded defects**, whose trace must be found by independent inquiry (for example, an undeclared import that the documentation does not mention). Detection of recorded defects shows that the methodological examination uses information that conventional checks do not examine; detection of unrecorded defects tests whether the examination can find what the documentation does not state. (4) Compare the conclusions drawn in each frame with the ground truth of Appendix A.2.

Measures. Differences in standard metrics between sound and compromised frames; guard detection rate and false-flag rate, reported separately for methodological (b, c) and computational (a, d) failures; inter-reviewer agreement on Adm_M ; error rate of conclusions in compromised frames relative to ground truth.

Interpretation of divergent results. If established auditing procedures detect some foundational defects, the result narrows the class of defects for which the guard is the sole safeguard and specifies its contribution precisely. If reviewer agreement on Adm_M is low, the result establishes that the documentation protocol requires greater precision; in line with Section 5.3, this bears on the implementation of the admissibility conditions, not on Law 3.

A.4. Experiment E2 — procedure

Conditions.

- (i) **Constructed semantic drift:** outputs modified toward or away from declared positions in the domain, at graded magnitudes, each validated by annotators.
- (ii) **Controls:** paraphrase, stylistic, and length changes confirmed by annotators as meaning-preserving.
- (iii) **Naturalistic collective drift:** successive self-training generations or successive model versions, with annotated samples.

Measures. Discrimination of semantic drift from meaning-preserving change (AUROC) for R1–R4; false alarms on controls; correlation between measured deviation and annotated magnitude; agreement between the axis to which R4 attributes drift and the direction identified by annotators (R4 and R3a; R3 provides no attribution); and, for condition (iii), the share of the common component, $\|\bar{\delta}\|$ relative to the mean $\|\delta_t\|$, in the full space and in projected coordinates ($\|U^\top \bar{\delta}\|$).

Interpretation of divergent results. If R4 and R3a perform similarly, the result locates the current advantage in the axis technique for this domain and encoder, and directs attention to the specification of

the foundational order in the frame documentation; the force of Law 4 against output-derived references remains established by Proposition 2. If the common component is small in the tested setting, Proposition 2 remains established, and the result specifies the drift regimes in which its consequences are most pronounced.

A.5. Experiment E3 — procedure

Frames. R4: full foundational order, documented. R3a: the same technique, number of axes, and number of formulations, from convenience examples. R4u: axes derived without supervision from in-domain data (principal components).

Ground truth. Domain experts judge the positions and pole memberships of held-out items. Independence is ensured by four rules: the experts take no part in writing *Def*, the pole formulations of R4, or the convenience examples of R3a; they are blind to which frame produced any axis; they judge items, not axes; and their judgments are collected before any frame is projected onto the held-out items.

Measures. Rank correlation between projected coordinates and expert judgments; discriminability on held-out items; separation of in-domain from out-of-domain items by the departure indicator r_t (AUROC).

Control. The matching criteria — number, length, and lexical diversity — are fixed before either frame is constructed. R3a examples are gathered by a separate team under a declared sampling procedure, without access to R4's poles; matching is verified afterwards by blind annotators, who also rate quality for both frames. The comparison thereby isolates the foundational order from expert curation without giving either frame an unintended advantage.

Interpretation of divergent results. If R4 exceeds R4u but not R3a, the result establishes that axes built from declared or selected examples outperform axes extracted from data in this setting; the additional effect of the foundational order beyond example selection then requires the distinguishing elements of the order to be made more explicit in the frame documentation. If R4 exceeds R3a but not R4u, data-extracted structure captures domain distinctions in this setting, and the comparison of R4 with R4u is examined by domain and by axis.

A.6. Experiment E4 — procedure

Concepts. Two candidate concepts whose operative definitions differ across centralities: *family*, continuing the family domain of Paper 02; and *progress*, whose meaning depends on the scale through which it is defined (Paper 03, Section 5.1). Each is retained only if *Rec* produces distinguishing elements sufficient for at least six distinguishing axes per centrality.

Frames. For each concept, two frames F_A and F_B , founded on two centralities, each recognized through *Rec* with documented source framing, derivation traces, and distinguishing elements. Both frames satisfy $\text{Adm}_M = 1$. The experiment describes the encoder relative to the two framings; it does not evaluate the framings.

Axes. Each frame contains at least six shared and six distinguishing axes, in equal numbers across the two frames. Each shared distinction is specified once, by a definition admitted under both centralities with the same pole content, and its formulations are written separately within each frame under the same protocol; blind annotators judge whether corresponding shared axes express the same distinction, and pairs below a pre-specified agreement threshold are excluded together with their counterparts. Distinguishing axes are derived from the distinguishing elements of each centrality. The minimum

design therefore comprises $m = 48$ axis instances per encoder (12 shared and 12 distinguishing instances per concept).

Matching and authorship control.

- Formulation sets are matched across frames in number per pole, length, lexical diversity, and register.
- Two independent writer groups, each competent in both framings, write for both frames, so that writer effects are distributed equally (counterbalanced authorship).
- Writer identities are recorded. The mode of production — independent or paired — is declared in advance for each concept. Each writer contributes equal numbers of formulations to both poles of every axis instance in which the writer participates; formulations by the same writer for the same pole satisfy $\text{dup} < \tau$; and no formulation is produced by modifying another formulation of the same pole.
- With $m = 48$ and $q = 0.05$, the feasibility conditions of Appendix A.7 give $B + 1 \geq 4,281$ and, for the writer-swap test, $m_w \geq 14$ writers per axis instance; $B = 10,000$ is adopted. Final writer numbers are set by power analysis.
- Blind annotators rate formulation quality, reported for each frame and axis type.

Encoders. Three encoders selected to differ in their training corpora — for example an English-centric encoder, a multilingual encoder, and an encoder with substantial Arabic-language training. Variation of the orientation across encoders with different corpora is itself part of the characterization of Γ_L , reported as covariation. Where available, a pair of encoders close in architecture, dimensionality, and objectives but differing in documented training data is added as a controlled comparison.

Measures.

- **Primary:** the orientation index $\mathcal{O}(\varphi; F_A, F_B)$ of Section 6.15, with the cell-level geometric means G , the realization proportions v_{sh} and v_{dist} , the number of collapsed axis instances and of pooled-floor values per cell, and the two sensitivity analyses of Appendix A.8.
- **Lexical control:** each separation is also computed with a lexical encoder (TF-IDF). After capping each separately, the lexically adjusted separation $L_k = \log \bar{S}_k^\varphi - \log \bar{S}_k^{\text{lex}}$ replaces $\log \bar{S}_k$ to give $\mathcal{O}_{\text{lex}}$. Because $\mathcal{O}$ is a double difference, constant offsets between the two encoders' scales cancel; non-linear differences and differences in how scale affects shared and distinguishing axes do not, and $\mathcal{O}_{\text{lex}}$ is therefore reported alongside $\mathcal{O}$ as orientation beyond that achieved by the specified lexical control.
- **Auxiliary:** γ_k and the departure contrast $\Delta \bar{r} = \bar{r}_H(F_B) - \bar{r}_H(F_A)$, reported separately and not combined into $\mathcal{O}$.
- **Robustness:** leave-one-axis-out recomputation of $\mathcal{O}$; split-half reliability of $\bar{S}_k$ within each pole.

Statistical analysis.

- *Primary estimate.* For each encoder and concept, $\bar{S}_k$ is computed for each axis instance from the pooled formulations of both writer groups, and $\mathcal{O}$ is computed from its definition; it is also reported for each writer group separately.
- *Uncertainty.* Writers contribute to several axis instances, so writers and distinctions form a crossed design. Confidence intervals for $\mathcal{O}$ are obtained by a two-way (pigeonhole) bootstrap (Owen, 2007), resampling at each replicate: (1) writers, once per replicate within each writer

group, each resampled writer carrying all of their formulations across all axis instances; (2) distinctions, independently — shared distinctions as pairs keeping their two frame instances together, distinguishing distinctions within their frame; (3) formulations within each writer, pole, and axis instance. Writer groups are not resampled.

- *Model-based confirmation.* A mixed-effects model is fitted with the formulation set (k, f, w) as the observational unit:

$$y_{kfw} = \beta_0 + \beta_F F + \beta_T T + \beta_{FT} (F \times T) + \beta_W W + d_k + a_{kf} + \epsilon_{kfw},$$

where $y_{kfw} = \log \bar{S}$ computed from the formulations of set (k, f, w) ; F , T , and W are fixed effects for frame, axis type, and writer group; d_k is a random intercept for distinction identity, shared by the two instances of a shared distinction; and a_{kf} is a random intercept for the axis instance. Formulations enter only through the computation of y_{kfw} . Because the logarithm of a pooled separation is not the mean of group-level log separations, β_{FT} and $\mathcal{O}$ are related but distinct estimands; $\mathcal{O}$ is primary, and agreement in sign and magnitude is reported as confirmation.

- *Scope of inference.* The unit of inference is the distinction. Formulations from different framings are never treated as exchangeable; exchangeability is assumed only for pole labels within a stratum, under the null hypothesis of the realization test.

Interpretation of divergent results.

- *No orientation* ($\mathcal{O} \approx 0$ with comparable separation levels in all four cells): the encoder realizes the distinguishing content of both framings comparably for these concepts. This characterizes Γ_L as not discriminating between the two framings on these concepts, and directs the selection of concepts with sharper distinguishing elements. It does not bear on the existence of Γ_L .
- *Distinguishing axes weak in both frames, shared axes strong* ($\mathcal{O} \approx 0$, established from the cell-level separations G and from v_{dist}): the encoder retains what the framings share and flattens what distinguishes them. This is itself a characterization of Γ_L .
- *Encoder collapsed for a concept* (established by the independent collapse condition): reported as a substantive characterization (Appendix A.8).
- *Index not computable under the floor rule*: reported as such, with no conclusion drawn about Γ_L .
- *Orientation varying with writer group*: the result reflects authorship rather than framing, and the matching protocol is refined before interpretation.

A.7. Realization test

Realization is assessed for each axis instance by a permutation test whose structure follows the declared mode of production of the formulations:

- *Independent production*: each formulation is written independently, and each writer contributes equal numbers of formulations to both poles of the axis instance. Pole labels are permuted within each writer, preserving the number of formulations per pole; strata are combined across both writer groups.
- *Paired production*: each positive formulation is written with a corresponding negative formulation, and labels are exchanged within pairs only.

In both modes the statistic is the raw separation $\bar{S}_k$ as defined above, with a_k , v_k , and $s_k^\pm$ recomputed under each permutation, since the direction is itself estimated from the labels. Under the null hypothesis that the encoder does not separate the poles, formulations are exchangeable between poles within each stratum. The protocol supports this assumption by requiring that formulations by the same writer for the same pole are not near-paraphrases of one another ($\text{dup} < \tau$) and that no formulation is produced by modifying another formulation of the same pole. Where within-writer dependence cannot be excluded, a coarser **writer-swap test** exchanges the complete pole sets of each writer as units.

An axis instance is **realized** if its separation is significant after Benjamini–Yekutieli control of the false discovery rate at a declared level q , across all axis instances of the same encoder (Benjamini & Yekutieli, 2001). Benjamini–Yekutieli is the primary procedure because it controls the false discovery rate under arbitrary dependence. Writers contribute to several axes, so the tests are dependent, and the condition under which Benjamini–Hochberg (1995) is guaranteed to be valid — positive regression dependency on the subset of true null hypotheses — cannot be established for these statistics; Benjamini–Hochberg results are reported as sensitivity analysis. With m axis instances in the test family and $H_m = \sum_{i=1}^m 1/i$, the number of permutations B must satisfy

$$\frac{1}{B+1} < \frac{q}{m H_m},$$

and the writer-swap test is admissible only if $2^{1-m_w} < q/(m H_m)$, where m_w is the number of writers contributing to each axis instance. These are feasibility conditions; sample sizes are fixed by power analysis.

v_{sh} and v_{dist} denote the proportions of realized shared and distinguishing axis instances. The realization classification is reported alongside the orientation index and does not enter it.

A.8. Floor rule, collapse, and dependence

Floor rule. The floor is used for one purpose only: to assign a value to an axis instance with $a_k = 0$. This value enters the realization profile and the orientation index as a diagnostic quantity; it does not turn an undefined axis into an operational one (Section 6.6). For each axis instance, every permuted value is evaluated by the same case table: a permuted value with $a_k \neq 0$ and zero dispersion is set to S_{max} ; one with $a_k = 0$ and positive dispersion is set to 0; one with $a_k = 0$ and zero dispersion is undefined and discarded. All values are capped at S_{max} . Then:

1. **Axis-specific floor.** If at least N_{null} valid null values remain and their median is positive, $S_{\text{floor},k}$ is that median: the separation expected from sampling variation alone.
2. **Pooled floor.** Otherwise the floor of that axis instance is non-estimable, and $S_{\text{floor},k}$ is set to the pooled floor: the median of the positive axis-specific null medians of the other axis instances of both frames, for the same encoder and concept, provided at least N_{pool} such medians exist. Because it is computed across both frames, the pooled floor confers no advantage on either.
3. **Index not computable.** If an axis instance with $a_k = 0$ needs a floor and fewer than N_{pool} axis-specific medians are available, $\mathcal{O}$ is reported as **not computable under this rule** for that encoder and concept. No empirical conclusion about I_L is drawn from this outcome.

Collapse as a separate finding. The number of collapsed axis instances, established by the independent condition above, is reported for each cell (frame $\times$ axis type). If most axis instances of a concept are collapsed and the encoding pipeline has been checked for technical faults (truncated, empty, or incorrectly batched inputs), the encoder is reported as **collapsed for the concept**: it does not distinguish

the formulations of either framing. This finding rests on the collapse condition itself, not on the behaviour of the null distribution.

Dominance of the pooled floor. The floor value replaces the observed separation only when $a_k = 0$. If the proportion of axis instances in any cell whose value is set by the pooled floor exceeds a declared ceiling λ_{pool} , $\mathcal{O}$ is reported as descriptive and flagged, and is not used as a primary result for that encoder and concept, since a common floor would then dominate the comparison.

Dependence structure. Two kinds of inference use different units. The realization test operates within a single axis instance, on formulations, with exchangeability only within a writer or a pair; dependence across axis instances, created by writers who contribute to several axes, is handled by the Benjamini–Yekutieli procedure. Confidence intervals for $\mathcal{O}$ treat writers and distinctions as crossed factors and use the two-way bootstrap of Appendix A.6; no formulation is treated as an independent observation in either procedure.

Two sensitivity analyses are reported: $S_{\text{floor},k}$ set to the lower quartile of the valid null values, with the same pooled fallback; and a chance-floored separation in which every $\bar{S}_k$ below its own $S_{\text{floor},k}$ is raised to that floor.

A.9. General statistical analysis

Expectations, thresholds, test families, and analysis plans are specified before data collection. Confidence intervals are obtained by bootstrap resampling matched to the structure of each experiment. Multiple-comparison corrections are applied across reference conditions. Sensitivity analyses vary ρ , ε , $\gamma_{\min}$, $S_{\min}$, $S_{\max}$, and the floor rule. Sample sizes are set by power analysis; the minimum material required to compute a statistic is not confused with the material sufficient to support inference. The values of all declared constants are fixed in an operational protocol before the experiments.