
ON MEASURING GROUNDING AND GENERALIZING GROUNDING PROBLEMS

PREPRINT

• Daniel Quigley
Center for Possible Minds
Indiana University Bloomington
Bloomington, IN 47408
dgquigle@iu.edu

Eric Maynard
Eruditis
Milwaukee, WI 53207
eric@eruditis.com

December 5, 2025

ABSTRACT

The symbol grounding problem asks how tokens like *cat* can be *about* cats, as opposed to mere shapes manipulated in a calculus. We recast grounding from a binary judgment into an audit across desiderata, each indexed by an evaluation tuple (context, meaning type, threat model, reference distribution): authenticity (mechanisms reside inside the agent and, for strong claims, were acquired through learning or evolution); preservation (atomic meanings remain intact); faithfulness, both correlational (realized meanings match intended ones) and etiological (internal mechanisms causally contribute to success); robustness (graceful degradation under declared perturbations); compositionality (the whole is built systematically from the parts). We apply this framework to four grounding modes (symbolic; referential; vectorial; relational) and three case studies: model-theoretic semantics achieves exact composition but lacks etiological warrant; large language models show correlational fit and local robustness for linguistic tasks, yet lack selection-for-success on world tasks without grounded interaction; human language meets the desiderata under strong authenticity through evolutionary and developmental acquisition. By operationalizing a philosophical inquiry about representation, we equip philosophers of science, computer scientists, linguists, and mathematicians with a common language and technical framework for systematic investigation of grounding and meaning.

Keywords symbol grounding · evaluation metrics · causal representation · compositionality · robustness · semantics · large language models · embodied AI

1 Introduction

The **symbol grounding problem** (SGP) asks how a token such as *cat* can be *about* cats, as opposed to being a shape manipulated in a formal system [30]; without such anchoring, definitions risk an infinite regress [75], in which each symbol is defined only in terms of other symbols, never making contact with the perceptual or sensorimotor experiences that could give them reference.

We use here “grounding” to mean an agent-internal, learned, task-indexed linkage from symbols to meanings that is accurate, causally earned, robust to declared disturbances, and systematic under composition. We frame grounding, then, as a check against criteria, and equip the discussion with variability; in this way, we move from a binary is-grounded and is-not-grounded, to an audit of *the extent to which* something is grounded. We operationalize this along four desiderata (preservation; faithfulness; robustness; compositionality), together with a zeroth, authenticity, which forbids post-hoc stipulation by requiring that the relevant semantic mechanisms reside within the agent, not supplied by external analysts. Our framework assigns checks at each desideratum; approximate grounding becomes appropriate, supplied with a common language.

Consequences follow: model-theoretic semantics meets structural criteria but lacks etiological warrant; text-only LLMs achieve partial grounding through distributional regularities of language; human language exemplifies comprehensive

grounding via multimodal perception and social coordination. Additionally, we gain in transparency about the behavior of understanding in a natural or artificial model when we expand from a single binary to a larger report, an insight into which targets should be prioritized for refinement and study when understanding is our subject matter, and a record against which we may match tasks such that we encourage trust in a model’s understanding. Finally, we provision a common language to discuss grounding problems in general, recovering the taxonomy discussed in [48].

Our enterprise is, therefore, provisioning a common language for grounding problems that operationalizes an inherently philosophical problem of meaning and representations of meaning. We equip philosophers of science, computer scientists, linguists, and mathematicians insight and a framework into the problem space with connections between empirical interpretability and benchmarking and foundational questions in semantics and cognitive science. In this way, we follow [29] in defining the problem space and operationalizing heretofore primarily philosophical concerns.

This paper is arranged as follows. Section 1 sets the introduction; Section 2 traces the symbol grounding problem from its philosophical roots in Searle’s Chinese Room and Harnad’s formulation to its modern consequences, showing how decades of internalist, externalist, and embodied approaches still leave grounding unresolved, and motivating our proposal. Section 3 introduces a neutral, diagnostic framework such that grounding claims become measurable profiles rather than metaphysical verdicts. Section 4 classifies how different architectures realize grounding within the general framework according to grounding modes as distinct ways systems link symbols, internal representations, and meaning spaces, each characterized by its own structural and causal constraints. Sections 5, 6, 7 evaluate model theoretic semantics, LLMs, and natural language according to their grounding profiles, respectively. Section 8 admits our implications, limitations, and biases in a forum of discussion, and Section 9 concludes. Section A proves theorems about our grounding framework, followed by Section B that walks through a simple toy example in grid-world. Section C provides a reference for notation used throughout.

2 Origins and implications

When [63] imagined a man shuffling Chinese symbols he cannot understand, the SGP followed; [30] formulated the parable as a research program: unless formal tokens connect to non-symbolic representations, infinite regress threatens the whole enterprise. Early debates contrasted agent-relative (internalist) with world-anchored (externalist) views [65]. Despite many proposals, “none provides a valid solution to the SGP, which remains open” [75].

The engineering corollary is clear: artificial agents whose internal symbols are grounded in their operational world. An agent instructed *Pick up the red mug* must link *red* to visual data and *mug* to object data [18, 13]. GOFAI’s hand-coded symbols were meaningful mainly to programmers; embodied robotics [11] argued for sensorimotor grounding; hybrids include Harnad’s iconic \rightarrow categorical \rightarrow symbolic pipeline, communication games [72], and multimodal neural agents and RL that tie reward to referential accuracy [33, 84]. A consensus solution to resolving grounding, however, remains elusive: approaches meet some aspects of grounding while neglecting others: our framework seeks to make explicit grounding according to such desiderata, and so turn a philosophical puzzle into a verifiable audit.

Let it not be said that we forgo accuracy for a suite other benchmarks; rather, accuracy is a useful number, but a poor story about understanding: it rewards getting the right label on a fixed set of items, and quietly assumes that the same trick will work when the wording, context, or nuisance factors change. That assumption is often false for modern AI [42, 47]; small paraphrases [38], typos [76, 44], or shifts in context [2] impact performance even when the original accuracy is high. In our terms, accuracy is, at best, an episode of correlational faithfulness under one specific evaluation tuple; it says nothing about whether the internal states were selected for success, whether meaning is stable under the threats that actually occur, or whether the system builds the whole from the parts in a principled way. Treating accuracy as proxy for understanding mistakes a point for the curve.

Grounding as evaluation replaces that single number with a profile. We declare the context and meaning type; we state the threat model and reference distribution, and audit basic preservation of atomic meanings, causal contribution, chart how semantic error grows with noise, and probe composition on held-out forms. Accuracy reappears inside this picture as a special case, when the semantic distance is mismatched¹; optimizing for that scalar alone invites Goodhart’s law [45, 32].

We might well ask whether accuracy can be made meaningful with better test design: yes, provided multiple parallel forms are built to enforce measurement invariance, if nuisance factors are systematically varied, if confidence intervals and item analyses are reported. In effect, that is our proposal: make the invariances and nuisances explicit; report the full profile; quote accuracy as one coordinate in a higher-dimensional audit. So, accuracy may be a red herring; it is not, however, that red herrings are inedible, only that a healthy diet requires other nutrients.

¹Technically, $1 - \text{accuracy}$ is our $\varepsilon_{\text{faith}}^{k,t}$ for that profile.

3 Generalized grounding

Let us provide now a common language for grounding.

Definition 3.1

Grounding is the degree to which an agent’s own mechanisms implement and have acquired mappings from symbols to the task-appropriate meaning space, such that realized meanings align with intended ones, that alignment was selected-for because it causes task success, the mappings remain stable under declared perturbations, and are systematic under composition.

The postulates that follow are diagnostic (indeed, we defer metaphysical commitments to further discussion elsewhere), and we remain neutral on whether meanings ultimately supervene on internal states or include wide-world factors [12, 57]. In this way, we build a framework from which we may predict behaviors and shape the discussion about grounding more generally, with later work given to application to models of interest. We itemize with the prefix “G”, for “grounding”.

Preservation (G1) atomic symbols preserve their intended meanings; encoding then decoding does not garble intent (beyond a small tolerance);

Faithfulness (G2) two parts:

1. *Correlational (G2a)* realized meanings match intended meanings (within tolerance);
2. *Etiological (G2b)* internal states were *selected for* success in the relevant task family;

Robustness (G3) given an explicit threat (e.g., typos, paraphrase, accent/ASR noise, lighting/pose), small input perturbations yield bounded semantic change such that grounded systems degrade gracefully;

Compositionality (G4) meaning is determined (up to tolerance) by the meanings of its parts, and the system generalizes to novel combinations with systematicity instead of relying on lookup for each new phrase.

We add a *zeroth* clause, **Authenticity**, demanding that the relevant mappings are implemented by the agent itself, not stitched on by an outside, omniscient analyst. We distinguish *G0-weak* (mechanisms implemented inside the agent) from *G0-strong* (implemented and acquired by agent-internal learning or evolution under process). External, post-hoc stipulation violates authenticity; it is cheating to offload and consult a translator when attempting to converse in Klingon, without knowing the language.

Example 3.1 (Kitchen example). *Consider the following: “Pick up the red mug to the left of the bowl”. We are interested to know whether and the extent to which such words and symbols connect to what is in front of the agent. How might we wonder at this?*

The context k is this particular kitchen: its layout; lighting; camera angle; who is speaking; when. The meaning type t is world-referential: does the phrase “the red mug to the left of the bowl” latch onto the correct object and relation in this scene? The threat U describes the kinds of small trouble we expect the agent to ride out: word-error in speech recognition; the typos or paraphrases; changes in lighting and pose; (mild) occlusions. The reference distribution P is the population of situations we judge against: new kitchens drawn from a held-out pool (a safety valve, such that we are not quietly grading on a curve).

Authenticity comes first (G0). In the weak form, the mechanisms that map words to internal states and then to concepts actually run inside the agent; the heavy lifting is not done for the agent off to the side by God, or more humbly, a human documentation, for example. In the strong form, the agent acquired those mechanisms through its own training or learning that was relevant to kitchen work; nothing crucial was bolted on after the fact. If semantics is smuggled in by the evaluator, the rest of the behaviors are misleading!

Preservation (G1) follows: do single words keep their meaning through the interpretation in this scene? When the agent hears `red`, `mug`, `bowl`, or `left_of`, the interpretation that emerges should still match what those meant here (within a small, declared error budget).

Faithfulness (G2) asks whether the system means what it seems to mean for the full command. There are two parts: first, correlational (G2a): for the whole sentence, does the interpreted meaning match what the speaker intended (again within a declared tolerance); second, etiological check (G2b): were the agent’s internal representations actually shaped by success on this family of tasks?

Robustness (G3) guards against meaning whiplash: when we inject perturbations from a threat, the agent’s interpreted meaning should change only a little, not jump to a wholly different object or relation. The intuitive picture is a smooth curve: as noise grows, error grows steadily and stays under a bound.

Compositionality (G4) is the promise that parts combine systematically. The meaning of a phrase should come from the meanings of its parts plus how they are put together. An agent that has learned “left of” and “behind” separately should cope with “the red mug behind the bowl” in a new kitchen layout without memorizing that exact sentence.

Let us consider: are these sufficient? Necessary? Indeed, there could be more criteria, or fewer. As it stands, what we present seems to be sufficient to direct the discussion to an operationalizing of grounding problems, and sufficient for walking through interesting consequences (Section 4, Sections 5, 6, 7, and Section A). We hope to encourage further discussion on this matter in future investigations.

Let us also clarify: grounding is *not* general intelligence; rather, it is task-relative, and audits how meanings are established and maintained, as opposed to what the system can achieve. Indeed, a model may be highly capable yet ungrounded for certain tasks, or narrowly capable yet well grounded within its domain; our desiderata capture the former, not the latter.

3.1 Formalization

Now formally, let us populate these clauses such that we may understand their behaviors. We wish to understand the shape of behavior; measurements with a ruler are exactly the next steps thereafter. Let K be a set of contexts (situations, tasks, Kaplanian contexts); let meaning types be $T = \{\text{ext}, \text{inf}, \text{soc}\}$ for *extensional* (world-referential), *inferential* (relational/intensional), and *social-normative* content. We assume that, for each pair of context and meaning (k, t) , there is an associated space \mathcal{M}_k^t with (pseudo)metric $d_{k,t}$; we set aside for now what this metric is in practice, and instead use it to illustrate and discuss behavior.

Σ is the agent’s observable “front end”: the alphabet of surface forms it consumes and produces. Depending on the architecture, these may be text tokens, phonemes, program tokens, action symbols, or other such items. We distinguish *atoms* $\Sigma_{\text{atom}} \subseteq \Sigma$ from composites built by a grammar \mathcal{G} . Typing is part of the surface discipline: constructors in \mathcal{G} only accept well-typed tuples from Σ . Note: technically, nothing in our framework forces Σ to be *linguistic*, though language is the primary motivation here; it is whatever the agent takes as manipulable symbols, objects, things.

\mathcal{R} is the set of agent-internal states that immediately follow encoding Φ . Formally, $\Phi : \Sigma \rightarrow \mathcal{R}$ (and, more generally, from upstream sensory/preprocessing streams into \mathcal{R}). We equip \mathcal{R} with a (pseudo)metric $d_{\mathcal{R}}$, such that we may talk about small versus large changes in representation, whatever those may turn out to be. Perturbations $u : \mathcal{R} \rightarrow \mathcal{R}$ for robustness act here (G3), and whose discreteness/continuity properties drive Corollary A.1.

\mathcal{C} is the agent’s task-level conceptual substrate. $\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ maps low-level representations to concepts, wherein we expect algebraic structure (types/constructors) to live when we assess compositionality (G4). Alignment $\mathcal{A}_k^t : \mathcal{C} \rightarrow \mathcal{M}_k^t$ projects concepts into the typed meaning space for the declared context and meaning type. Only the composite $S_{k,t} := \mathcal{A}_k^t \circ \Gamma$ is observable; \mathcal{C} is not unique (see the quotient construction below), and changes to \mathcal{C} that leave $S_{k,t}$ invariant are a kind of “gauge” transformations with no semantic effect. Interestingly, we remain agnostic to what this \mathcal{C} looks like, only that we need it for passing from one space to another, as we shall see below.

For each context $k \in K$ and meaning type $t \in T = \{\text{ext}, \text{inf}, \text{soc}\}$, \mathcal{M}_k^t is the typed domain of meanings against which we evaluate the agent. We equip \mathcal{M}_k^t with a (pseudo)metric $d_{k,t}$ and, when relevant, an algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ interpreting the grammar’s typed constructors. The intended interpretation $\mathcal{I}_k^t : \Sigma \rightarrow \mathcal{M}_k^t$ provides meanings for atoms (and, via its homomorphic extension, for composites).

Let us illustrate these with some cartoon examples to build the intuition.

Example 3.2 (Vision task). For $t = \text{ext}$ in a vision-manipulation task (see [68, 27]), \mathcal{R} a stack of convolutional/-transformer features with cosine $d_{\mathcal{R}}$; \mathcal{C} encodes object candidates and spatial relations with a graph metric $d_{\mathcal{C}}$; $\mathcal{M}_k^{\text{ext}}$ consists of scene-anchored objects and relations with $d_{k,t}$ defined by $1 - \text{IoU}$ plus relation penalties.

Example 3.3 (Logical system). Relatedly, for $t = \text{inf}$, $\mathcal{M}_k^{\text{inf}}$ a logical or relational structure with operations \mathcal{F}_k^t interpreting connectives and predicates; $d_{k,t}$ entailment-based or graph-edit distance (see [62, 39, 67]).

Example 3.4 (Social setting). For $t = \text{soc}$, we remain abstract, for such measurements are difficult to determine, but we play the same game to illustrate the point: $\mathcal{M}_k^{\text{soc}}$ dialogue commitments or normative statuses with $d_{k,t}$ tied to expected repair cost (see [21]).

For each context-type pair (k, t) , $\mathcal{I}_k^t : \Sigma \rightarrow \mathcal{M}_k^t$ partially specifies “gold” meanings for atoms in that setting². When a semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ is declared, \mathcal{I}_k^t extends uniquely to composites as the homomorphic map $\mathcal{I}_k^{t, \uparrow}$ (Corollary A.2). Methodologically, \mathcal{I}_k^t fixes the target against which preservation and faithfulness (G1–G2) are measured; it does *not* by itself endow the agent with meaning (regulated by G0).

$\Phi : \Sigma \rightarrow \mathcal{R}$ maps surface symbols to internal representations, whatever those may be. In text systems, for example, this is simply a tokenizer plus embedding; in speech, ASR front ends; in robotics, perception that convert symbolic commands to latent states. We treat Φ as the causal “portal” from symbols to the representation space \mathcal{R} equipped with metric $d_{\mathcal{R}}$. Under G0-strong, Φ is acquired by the agent for the task family under evaluation; under G0-weak, it is *at least* implemented inside the agent.

$\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ turns low-level representations into task-level concepts in \mathcal{C} (metric $d_{\mathcal{C}}$). We need to account, for example, parsing, attention, relational inference, and other structure-building operations. Γ is where typed combinators are to be realized internally, which is why compositionality tests (G4) apply naturally *after* Γ . Only the composite to meanings, $S_{k,t} = \mathcal{A}_k^t \circ \Gamma$, is observable in evaluation³.

$\Psi = \Gamma \circ \Phi : \Sigma \rightarrow \mathcal{C}$ is the end-to-end map from surface symbols to concepts. It is technically *context-agnostic*: we keep contextual and meaning-type variation out of Ψ and place it in the alignment \mathcal{A}_k^t and the semantic algebra on \mathcal{M}_k^t . This is necessary in order to walk through preservation (G1) and correlational faithfulness (G2a), which compare $\mathcal{A}_k^t \Psi$ to \mathcal{I}_k^t at the chosen (k, t) ; etiological faithfulness (G2b) asks whether states along Ψ were selected for success.

For each (k, t) , $\mathcal{A}_k^t : \mathcal{C} \rightarrow \mathcal{M}_k^t$ projects the agent’s concepts into the typed meaning space relevant to that evaluation. This is how we track context along with the semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$: scene particulars; speaker; time; task constraints; meaning type. Robustness (G3) checks the continuity of $S_{k,t} = \mathcal{A}_k^t \circ \Gamma$; compositionality (G4) checks whether $\mathcal{A}_k^t \Psi$ respects the constructors’ interpretations $f_k^{\mathcal{M}, t}$. Under G0-strong, \mathcal{A}_k^t is an agent-internal, learned mechanism (e.g., learned perceptioworld alignment); if \mathcal{A}_k^t is an external measurement adapter, then claims must be G0-weak to avoid smuggling semantics. Laconically, \mathcal{A}_k^t is the “semantics”.

Definition 3.2

A **grounding architecture** is

$$\mathfrak{G} = \langle \Sigma, \mathcal{R}, \mathcal{C}, \{\mathcal{M}_k^t\}_{k,t}, \{\mathcal{I}_k^t\}, \Phi, \Gamma, \Psi, \{\mathcal{A}_k^t\}_{k,t} \rangle$$

We check the extent to which a grounding exists if preservation, faithful representation, robustness, and compositionality hold, subject to a zeroth postulate of authenticity. We distinguish:

G0-weak Φ, Γ are implemented inside the agent by causal mechanisms;

G0-strong as above, and moreover Φ, Γ and any used \mathcal{A}_k^t were *acquired* by agent-internal learning or evolution under a process \mathcal{T} relevant to the evaluated tasks; post-hoc stipulation violates G0.

We fix an evaluation $E = (k, t, U, P)$ with perturbations U and reference distribution P over \mathcal{R} (or upstream).

G1 (preservation) for atoms $\sigma \in \Sigma_{\text{atom}}$ and a chosen (k, t) ,

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t}.$$

Note: $\varepsilon_{\text{pres}}^{k,t}$ is estimated on an atomic set (words, idioms), *not* on composed items.

G2a (faithfulness, correlational) For (possibly composed) σ ,

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{faith}}^{k,t}.$$

Note: this is distinct from G1, which *does* permit composed items, and by (k, t) choice.

G2b (faithfulness, etiological) fix an evaluation $E = (k, t, U, P)$ and a success predicate $\text{succ} : \mathcal{O} \rightarrow \{0, 1\}$. We require two things about some *agent-internal* mechanism M (within Φ, Γ , and/or \mathcal{A}_k^t):

1. M was acquired under a declared training/evolutionary process \mathcal{T} because it improved succ on the task family;

²We require partiality to respect lexical gaps, ambiguity, or items outside scope, for exmaple.

³Note that different choices of Γ that induce the same $S_{k,t}$ are indistinguishable by our metrics

2. in the evaluated regime, turning M off would reduce success by at least a stated margin.

Formally, define the average causal effect, imaginatively noted as ACE:

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))],$$

with the expectation over the declared instance distribution (and, if relevant, over \mathcal{T}). We say G2b holds at level $\eta^{k,t}$ if $\text{ACE}_E(M) \geq \eta^{k,t}$ and M was produced by the declared \mathcal{T} rather than stipulated post hoc (checked by G0).

G3 (robustness) for perturbations $u : \mathcal{R} \rightarrow \mathcal{R}$ (e.g., typos, paraphrase, accent, noise, lighting change) with scale $\varepsilon(u)$ and confidence level $1 - \alpha$,

$$\Pr_{r \sim P} \left[\sup_{\varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^t \Gamma(r), \mathcal{A}_k^t \Gamma(u \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

with $\omega_U^{k,t}(0) = 0$. Note that Lipschitz continuity is the special case $\omega_U^{k,t}(\varepsilon) \leq L^{k,t}\varepsilon$.

G4 (compositionality) let the grammar provide typed constructors $f : \Sigma^{\bar{\tau}} \rightarrow \Sigma^{\tau'}$ and the semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ assign $f_k^{\mathcal{M},t} \in \mathcal{F}_k^t$. Then the grounding mode is $\delta_{\text{comp}}^{k,t}$ -compositional if

$$d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), f_k^{\mathcal{M},t}(\mathcal{A}_k^t \Psi(\sigma_1), \dots)) \leq \delta_{\text{comp}}^{k,t}.$$

Systematicity on a heldout set $\mathcal{D}_{\text{novel}}$ with tolerance τ is given by

$$\beta^{k,t} = \frac{1}{|\mathcal{D}_{\text{novel}}|} \sum_{(\vec{\sigma}, y) \in \mathcal{D}_{\text{novel}}} \mathbf{1}[d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), y) \leq \tau].$$

The discussion of these items is directed by scales, values, and tolerances; these are what we use to predict behaviors. Let us now tour and exposit upon these in greater detail.

3.2 G0 (authenticity)

G0 anchors the entire framework by stipulating that the semantic work must occur inside the agent rather than being supplied externally. In this way, we introduce a distinction between having mechanisms that operate within the system and having those mechanisms as the product of the system’s own learning or evolution. Without this authenticity requirement, meanings could be imported, declared, or maintained by an outside observer when we want them realized by the agent itself.

Under G0-weak, the agent’s encoding and conceptualization (its mappings Φ and Γ) are implemented inside the system and causally operative in producing behavior, but their contents may have been fixed by an external designer or analyst. Under G0-strong, these same mappings, together with any alignment \mathcal{A}_k^t used for evaluation, were acquired by the agent itself through an internal process of learning or evolution, denoted \mathcal{T} , that is relevant to the tasks being evaluated. Post-hoc assignment of meaning, where an observer or developer stipulates what the agent’s symbols stand for, violates authenticity. Meaning must arise through mechanisms that the agent not only possesses but has earned through its own causal history of adaptation or success.

The idea here is that authenticity separates an interpreted system from a self-interpreting one. A dictionary or a translation engine, for instance, certainly can map between words with high fidelity, yet neither possesses internal understanding of what those words denote. Using translate to render “cat” into Chinese and back again will reliably produce the same character sequence, but the translation system has not formed any concept of a cat: it manipulates patterns of symbols according to correlations extracted from data. This is Searle’s Chinese Room: the person inside the room follows an English rulebook that maps Chinese characters to appropriate outputs; to outside observers, the system behaves as though it understands Chinese, but all the semantic work is done by the rulebook’s author, not by the person or the room. The mechanism is internal in a physical sense but not authentic in a semantic one: the meaning has been smuggled in from outside.

Under G0-strong, the rulebook itself must have been learned or evolved within the agent. A human language learner satisfies this condition: whatever mechanisms that connect words to perceptions and actions are the result of developmental processes shaped by feedback from the environment and community. Similarly, an embodied agent that learns through reinforcement to associate the word “cup” with graspable cylindrical objects achieves authenticity if its perceptual and linguistic pathways were acquired through training that improved task success; labels hard-coded by designers are certainly useful, but now authentic accordingly. In contrast, a symbolic planner endowed with a

hand-written lexicon of predicates and definitions is at best G0-weak; its symbols are causally effective but semantically inherited.

The purpose of G0 is, therefore, methodological as it is philosophical, in which we enforce a control on what counts as evidence of grounding by distinguishing internal causal contribution from external interpretation. Once authenticity is established, the remaining desiderata follow: their measurements refer to mechanisms that the agent genuinely possesses. Without G0, the entire enterprise risks circularity: we could always make any ungrounded system *appear* grounded by attaching meanings from the outside; G0 compels us to locate meaning inside the agent’s own causal and developmental structure.

3.3 G1 (preservation)

The simplest grounding desideratum: the meanings of atomic symbols remain stable through the agent’s processing. Before a system can be said to combine, infer, or act on meanings, it must first preserve the integrity of the basic semantic units it operates upon. In this sense, preservation acts as a kind of “semantic conservation law” of grounding: if the meaning of “red” drifts toward “orange”, or “cat” begins to encode something closer to “dog”, then the subsequent faithfulness (G2), robustness (G3), or compositionality (G4) are already compromised.

Formally, for each atomic symbol $\sigma \in \Sigma_{\text{atom}}$ and a declared evaluation (k, t) , preservation holds when

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t}.$$

Here, $\Psi = \Gamma \circ \Phi$ is the system’s end-to-end map from surface symbols to its internal conceptual representations, \mathcal{A}_k^t aligns those internal concepts to the meaning space appropriate to the chosen context k and meaning type t , and \mathcal{I}_k^t provides the intended interpretation of the same atoms. The metric $d_{k,t}$ measures semantic distance in that meaning space, and $\varepsilon_{\text{pres}}^{k,t}$ is the maximal admissible deviation, an (empirically estimated) tolerance that specifies how much drift is acceptable before preservation is considered to have failed.

Consider: G1 asks whether the system’s internal realization of each basic term still means what it is supposed to mean within the declared context; it is, in effect, a calibration test: the system’s elementary symbols should behave as stable semantic units. Without this guarantee, higher-order semantic structure cannot be assessed, since every compositional or causal claim would rest on malleable, moving referents.

The metric $d_{k,t}$ adapts to the meaning type. In a vectorial or distributional setting, the cosine distance between the model’s embedding of a token and a “gold” embedding derived from human similarity judgments. In referential or perceptual tasks, measure spatial or perceptual overlap, such as $1 - \text{IoU}$ between predicted and true object regions for “cat” or “mug”. In logical systems, $d_{k,t}$ may very well collapse to identity: zero when denotations coincide, one otherwise. In social-normative contexts, this metric is more difficult to decide, for which here we posit a kind of “cost” of conversational repair or divergence in normative commitments. Whatever its form, $d_{k,t}$ checks for small versus large semantic discrepancy in the relevant domain.

Preservation underwrites the rest of the framework. Faithfulness (G2) presupposes that atomic meanings are correct before their combinations can be tested; robustness (G3) measures continuity relative to those preserved anchors; compositionality (G4) assumes stable building blocks for systematic combination. G1 thus secures the semantic “ground floor”, so to say, on which the higher desiderata rest.

Example 3.5 (Symbolic logic system). *For a symbolic calculus with atomic predicates $\Sigma_{\text{atom}} = \text{bachelor, man, unmarried}$, and an interpretation \mathcal{I}_k^t that stipulates their denotations, preservation holds exactly: $d_{k,t} = 0$ whenever $\mathcal{A}_k^t \Psi(\sigma) = \mathcal{I}_k^t(\sigma)$. Here $\varepsilon_{\text{pres}}^{k,t} = 0$; the atoms are preserved by definition.*

Example 3.6 (Large language model). *For an LLM, Σ_{atom} consists of subword tokens. The realized meaning $\mathcal{A}_k^t \Psi(\text{cat})$ follows from the contextual embedding of “cat”, while $\mathcal{I}_k^t(\text{cat})$ is a reference vector derived from human similarity judgments or dictionary definitions against which to check. Preservation error $\varepsilon_{\text{pres}}^{k,t}$ records how consistently the model keeps “cat” semantically near related words (“feline”, “kitten”) and distant from unrelated ones (“dinosaur”, “idea”). Polysemy and contextual drift will inevitably raise this error.*

Example 3.7 (Embodied agent). *For an agent interpreting “red”, “blue”, and “mug”, preservation tests whether its perceptual pipeline maps these terms to stable visual categories. Mis-identifying red mugs under new lighting would increase $d_{k,t}$ and reveal that preservation fails in that region of the perceptual space.*

Preservation therefore enforces the identity principle of grounding: if a symbol σ means m (by whatever representative formulation we take for “mean”), then the agent’s internal realization of σ should approximate m within a tolerance. Only once this semantic conservation is established can one meaningfully evaluate whether the system is faithful, robust, or compositional in its use of those preserved meanings.

3.4 G2 (faithfulness)

An agent’s realized meanings should match the intended ones, and those matches produced by mechanisms that were selected for success rather than stipulated or accidental. As such, it has a correlational tier (G2a) and an etiological tier (G2b), together distinguishing *what* the system does from *why* it does it.

For a declared evaluation setup (k, t) and any well-typed expression σ (atomic or composed, for now we allow for it), correlational faithfulness holds when

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{faith}}^{k,t}.$$

We have that $\Psi = \Gamma \circ \Phi$ is the agent’s end-to-end symbol-to-concept map, \mathcal{A}_k^t aligns those concepts to the meaning space \mathcal{M}_k^t relevant to context k and meaning type t , \mathcal{I}_k^t provides the intended interpretation against which we compare, $d_{k,t}$ measures semantic discrepancy in \mathcal{M}_k^t , and $\varepsilon_{\text{faith}}^{k,t}$ is a tolerance estimated that includes composed items. Unlike G1, which probes atoms and establishes a baseline, G2a evaluates whole expressions under the chosen (k, t) : the agent’s realized meaning should land within a tolerance of the target meaning. In practice, $d_{k,t}$ inherits the semantics of the task: in referential settings it may be $1 - \text{IoU}$ on denoted objects or a spatial-relation penalty; in vectorial settings, cosine distance between predicted and reference representations; in logical settings, identity of denotations or overlap of entailment sets. Low $\varepsilon_{\text{faith}}^{k,t}$ indicates that the agent behaves as if it understands, but by itself says nothing about causal responsibility.

To catch causality, then, (or, rather, attempt to; causality is a difficult problem to solve, to say the least) we consider etiological faithfulness. Fix an evaluation tuple $E = (k, t, U, P)$ and a success predicate $\text{succ} : \mathcal{O} \rightarrow 0, 1$ for the task family at issue. There must exist an agent-internal mechanism M (be it a learned subnetwork, circuit, module, rule, or parameter subset within Φ, Γ , and/or \mathcal{A}_k^t) that was acquired under a declared training or evolutionary process \mathcal{T} , because it improved success on that family, and whose current contribution is measurably causal. We require a positive average causal effect

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t},$$

with the expectation taken over the instance distribution in E (and, when relevant, over \mathcal{T}). The threshold $\eta^{k,t}$ sets sufficiency appropriate to the domain. An implementation is direct intervention or ablation that toggles M and demonstrates the requisite drop; when interventions are infeasible, lower bounds obtained via retrain-without- M controls, invariance across environments, instrumental-variable analyses, or influence-function approximations suffice, provided assumptions and uncertainty are stated. High accuracy alone does not satisfy G2b unless one links it, causally and historically, to such an M selected for success.

G2a asks whether realized meanings “line up” with intended meanings; G2b asks whether that alignment is “earned” by mechanisms the agent acquired because they made it succeed. One can satisfy G2a with a lookup table or brittle shortcut; G2b blocks such semantic smuggling by insisting on an internal, learned basis for success. Conversely, a mechanism might be causally potent yet misaligned with the evaluator’s targets; then G2a fails even if G2b holds for some other outcome.

In estimation, G2a reduces to semantic error on composed items under (k, t) ; one reports $\varepsilon_{\text{faith}}^{k,t}$ with confidence intervals, specifying the metric $d_{k,t}$, the composition of the test set, and any nuisance variation that has been controlled. G2b requires more care, however, in experimental design. Direct ablations of M , retraining without the signal that produces M , and environment interventions that change task-relevant covariates while holding spurious ones fixed, all contribute evidence toward $\text{ACE}_E(M)$. Because $\eta^{k,t}$ is task-dependent, etiological sufficiency is indexed to the evaluation tuple: an M that is selected-for linguistic prediction may meet G2b for $(k_{\text{ling}}, \text{inf})$ but not for $(k_{\text{embodied}}, \text{ext})$, for example.

Example 3.8 (Text-only language model). For $(k, t) = (k_{\text{ling}}, \text{inf})$, G2a measures how closely the model’s realized meanings match linguistic targets such as paraphrase semantics or entailment labels; $d_{k,t}$ a cosine loss or an entailment discrepancy, and $\varepsilon_{\text{faith}}^{k,t}$ will be small on in-distribution text.

For G2b, we have a difficult problem indeed, in identifying an internal head or subspace M that supports, say, coreference resolution, show that it emerged under the language-modeling objective \mathcal{T} , and that ablating M reduces success on coreference tasks by at least $\eta^{k,t}$. Such an exercise, though easy to state, is decidedly not so easy to practice, and certifies etiological faithfulness for linguistic success; it does not by itself establish G2b for world-referential tasks.

Example 3.9 (Embodied referential agent). For $(k, t) = (k_{\text{embodied}}, \text{ext})$, G2a compares denotations or spatial relations derived from full commands (e.g., “the red mug left of the bowl”) against scene ground truth; $d_{k,t}$ must then combine $1 - \text{IoU}$ with relation penalties.

For G2b, a learned visual-grounding module M that arose through reinforcement or imitation learning \mathcal{T} and whose ablation causes a marked drop in task success provides the required causal warrant. Here both alignment and selection-for are about world contact.

Example 3.10 (Logical knowledge base). In a symbolic system evaluated at $(k_{\text{logic}}, \text{inf})$, G2a holds trivially when $\mathcal{A}_k^t \Psi$ is defined to coincide with the stipulated \mathcal{I}_k^t ; then $\varepsilon_{\text{faith}}^{k,t} = 0$ on the covered fragment.

Unless the rules and interpretations were themselves acquired under a process \mathcal{T} that selects for inferential success, however, G2b fails right out: the mapping is correct by fiat, not earned by the agent.

G1 ensures that atomic meanings are intact; G2a extends agreement to whole expressions under the declared semantics; G2b ties that agreement to the agent’s causality and history.

3.5 G3 (robustness)

A grounded system must sustain the stability of its meanings under perturbation; it demands that small changes to input or context (typos; paraphrases; variations in lighting or viewpoint; shifts in register or accent) induce (proportionally) small changes in the system’s interpreted meanings. Robustness, in this sense, is the requirement of semantic continuity under disturbance: a grounded agent should not lose its grip on meaning when the world or its sensory inputs wobble.

For perturbations $u : \mathcal{R} \rightarrow \mathcal{R}$ drawn from a declared threat U with scale $\varepsilon(u)$ and confidence level $1 - \alpha$, robustness holds when

$$\Pr_{r \sim P} \left[\sup_{\varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^t \Gamma(r), \mathcal{A}_k^t \Gamma(u! \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

with $\omega_U^{k,t}(0) = 0$. Here, r denotes an internal representation drawn from the reference distribution P ; $u \cdot r$ is its perturbed counterpart; $d_{k,t}$ measures semantic distance in the meaning space appropriately relevant to context k and meaning type t ; Γ converts representations to conceptual states; \mathcal{A}_k^t projects those concepts into the space of meanings; and $\omega_U^{k,t}(\varepsilon)$ bounds the maximum tolerated drift under perturbations of scale⁴ up to ε . The confidence level $1 - \alpha$ expresses that the condition holds for the majority of representative inputs, rather than universally.

We want to account for “graceful degradation”. The robustness modulus $\omega_U^{k,t}$ acts as a semantic analogue of a shock absorber: it describes how fast meanings move when inputs are jostled. A well-grounded system should display a smooth, shallow curve around $\varepsilon = 0$: semantics change slowly under realistic disturbances; an ill-grounded or brittle system exhibits step-function behavior, in that it is unchanged until a tiny perturbation causes catastrophic semantic collapse.

To interpret this concretely, we must specify both the perturbation space and the metric of semantic deviation.

Example 3.11 (Linguistic system). In linguistic systems, the threat model U_{ling} may be quite variable, consisting of paraphrases, synonyms, typographical errors, with $\varepsilon(u)$ corresponding to edit distance or embedding-level noise. The internal representations r are text embeddings, and the semantic distance $d_{k,t}$ measured in the usual way by cosine difference between meanings, or by shifts in entailment probability. A robust language model should maintain nearly constant $\mathcal{A}_k^t \Gamma(r)$ under such surface perturbations: “the red mug” and “red mug” should denote the same entity.

Example 3.12 (Vision model). In vision and robotics, U_{nat} lighting variation, camera rotation, blur, occlusion. Here r denotes visual features, $\varepsilon(u)$ quantifies photometric or geometric change, and $d_{k,t}$ measures deviation in predicted scene meanings. A grounded agent maintains stable reference: the “red mug” remains the same object under mild viewpoint shifts.

Example 3.13 (Multimodal systems). For multimodal systems (U_{vl}), we expect that perturbations couple linguistic and visual channels, such as cropping paired images or paraphrasing captions. $d_{k,t}$ cosine distance in a shared embedding space or mismatch rate in retrieval. The local slope $L = \omega_U^{k,t}(0)$ then captures how quickly semantic alignment deteriorates under incremental corruption.

Example 3.14 (Multimodal systems). Symbolic and logical systems behave differently. Their representation space \mathcal{R} is uniformly discrete, and U_{edit} comprises single-character or rule edits. Because there are no intermediate states, $\omega_U^{k,t}$ degenerates: it is zero for $\varepsilon < 1$ and jumps abruptly when an edit occurs. Symbolic systems are therefore formally continuous, practically brittle: continuity holds only vacuously, since any small perturbation breaks the grammar or change meaning entirely.

Example 3.15 (Social settings). In social-normative or pragmatic domains, robustness tests the stability of communicative meaning under minor social variation, which may well include the above perturbations as well. U_{soc} register

⁴Lipschitz continuity is a special case where the modulus grows linearly, $\omega_U^{k,t}(\varepsilon) \leq L^{k,t} \varepsilon$.

changes or politeness shifts; r dialogue states; $d_{k,t}$ we take as the cost of conversational repair required to restore mutual understanding. Minor differences such as “thanks” versus “cheers” should produce low deviation, while an insult constitutes a large ε .

We sketch the empirical recipe: estimation of robustness proceeds by defining the threat model U , sampling unperturbed representations $r \sim P$, applying controlled perturbations $u \cdot r$, computing semantic deviations $d_{k,t}(\mathcal{A}_k^t \Gamma(r), \mathcal{A}_k^t \Gamma(u \cdot r))$, and fitting the empirical modulus $\omega_U^{k,t}(\varepsilon)$. Robustness thus quantifies the continuity of meaning with respect to both internal and environmental noise; preservation secures what the atoms mean, faithfulness ensures they mean the right things, and robustness ensures they continue to mean them under the minor degradations inevitable in real-world conditions, in which systems behave as though equipped with appropriate shock absorbers.

3.6 G4 (compositionality)

A grounded system should build the meanings of complex expressions systematically from the meanings of their parts; it operationalizes the classical principle of compositionality: the meaning of the whole is determined by the meanings of the parts and their mode of combination. In a grounded architecture, compositionality ensures that once the atomic meanings are preserved (G1) and correctly aligned (G2), they can be combined productively without retraining on every new phrase or configuration.

Formally, let the grammar provide typed constructors $f : \Sigma^{\vec{\tau}} \rightarrow \Sigma^{\tau'}$ that specify how symbols may combine, and let the semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ assign to each syntactic constructor f a semantic operation $f_k^{\mathcal{M},t} \in \mathcal{F}_k^t$. The system is $\delta_{\text{comp}}^{k,t}$ -compositional if, for every such constructor and tuple of arguments $\vec{\sigma} = (\sigma_1, \dots, \sigma_n)$,

$$d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), f_k^{\mathcal{M},t}(\mathcal{A}_k^t \Psi(\sigma_1), \dots, \mathcal{A}_k^t \Psi(\sigma_n))) \leq \delta_{\text{comp}}^{k,t}.$$

Here, $\Psi = \Gamma \circ \Phi$ maps surface symbols to internal concepts, \mathcal{A}_k^t aligns those concepts to the meaning space appropriate to the declared context k and meaning type t , $d_{k,t}$ measures semantic distance in that space, and $\delta_{\text{comp}}^{k,t}$ quantifies the tolerated deviation between the meaning that the agent actually realizes for the composed expression and the meaning obtained by composing the meanings of its constituents via the operations defined in the semantic algebra. When $\delta_{\text{comp}}^{k,t} = 0$, the mapping $\mathcal{A}_k^t \Psi$ is a homomorphism from the syntactic algebra to the semantic one: the system realizes perfect compositionality on its grammar (see Theorem A.2). Nonzero but small $\delta_{\text{comp}}^{k,t}$ indicates approximate compositionality, a deviation.

Systematicity on a held-out set of novel combinations $\mathcal{D}_{\text{novel}}$ further checks whether this compositional behavior generalizes. Given a tolerance τ , the systematicity

$$\beta^{k,t} = \frac{1}{|\mathcal{D}_{\text{novel}}|} \sum_{(\vec{\sigma}, y) \in \mathcal{D}_{\text{novel}}} \mathbf{1}[d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), y) \leq \tau]$$

reports the proportion of unseen expressions whose realized meanings remain within the acceptable distance τ of their target meanings; high $\beta^{k,t}$ is the ability to extrapolate compositionally as opposed to memorizing configurations.

Let us gesture at the metric used here as we have above. The distance metric $d_{k,t}$ checks for what counts as a compositional deviation, and varies with the meaning type⁵. In vectorial or distributional models, $d_{k,t}$ is geometric in the space, measuring how closely the representation of a composite phrase matches the vector predicted by composing its parts. In symbolic or logical systems, $d_{k,t}$ collapses to equality or an edit distance on derivations, probing whether the inferred logical form yields the same entailments as the algebraic combination of constituents. In referential and perceptual tasks, $d_{k,t}$ is spatial, again such as $1 - \text{IoU}$ for compared object regions, or an L2 distance between predicted and actual spatial relations. In social or pragmatic domains, we have that $d_{k,t}$ checks distance in a network of dialogue commitments or is simply the repair cost needed to restore mutual understanding; compositionality here measures whether the pragmatic force of a compound utterance derives predictably from its parts.

Whatever the metric, the goal is to quantify how far apart the composed meaning and the algebraic composition of its parts lie in the space where success or truth is judged.

⁵It is a fair question whether such metrics can simply be added with appropriate weights for meanings that span more than one domain, such as mixed or hybrid systems, and whether they should employ some composite metrics, for example

$$d_{k,t}(m_1, m_2) = \alpha d_{\text{ext}} + \beta d_{\text{inf}} + \gamma d_{\text{soc}},$$

to balance extensional, inferential, and social-normative components, for example. We leave this as an aside for now, and for future work.

Example 3.16 (Vectorial composition). *In a language model, embeddings for “red”, “car”, and “red car” can be compared. If the direct embedding of “red car” lies close to the vector produced by adding or linearly combining the embeddings for “red” and “car”, the model exhibits low $\delta_{comp}^{k,t}$: it composes meanings geometrically in a way that mirrors the language’s own algebra of modifiers and nouns.*

Example 3.17 (Referential composition). *For an embodied agent interpreting “the red mug left of the bowl”, the grammar provides a constructor combining color, object, and spatial relation. The semantic algebra interprets this as an intersection of color and object predicates constrained by a spatial relation.*

Compositionality requires that the meaning assigned by the agent to the full phrase match, within admitted tolerance $\delta_{comp}^{k,t}$, the intersection of its meanings for “red mug” and “left of the bowl”. If performance on unseen combinations like “blue bowl right of the mug” remains stable, the agent demonstrates high systematicity $\beta^{k,t}$.

Compositionality connects the local stability of meanings secured by G1–G3 to the productive structure of language and reasoning. G1 guarantees that atoms mean what they should; G2 ensures that those meanings are correctly aligned and causally warranted; G3 preserves them under perturbation; G4 ensures that they combine coherently to produce new meanings.

4 Varieties of grounding

Different $\{\Sigma, \mathcal{R}, \mathcal{M}\}$ instantiations yield different grounding modes, and so \mathfrak{G} subsumes grounding types discussed in [30, 28, 48, 71]. The grounding mode is simply a label for *how* an architecture (attempts to) link(s) symbols to meanings. Note that this is a property an architecture may satisfy (e.g., symbolic, referential, vectorial, relational), and is not the architecture itself. Framed in this way, we recover the taxonomy explored as well in [48], and provision for it an evaluation in our grounding framework.

Definition 4.1

A **grounding mode** at (k, t) is a predicate $\text{Mode}(\text{GA})$ over grounding architectures specified by structural and etiological constraints on \mathfrak{G} . We write $\text{GA} \in \text{Mode}$ when those constraints hold (e.g., *symbolic, referential, vectorial, relational*).

There exists a subtlety here. Our modes can be read on two axes: symbolic, vectorial, and relational modes describe how content is internally represented through rules, geometry, or structured roles; referential grounding cuts across these as the case where such representations are additionally embodied or sensorimotorily coupled to the world they describe. A system may therefore be symbolically, vectorially, or relationally referential, depending on how its internal substrate connects to its perceptual-action loop. Multimodality alone does not guarantee referential grounding: what matters is the causal provenance of data and the world-coupling of the learning process.

In this way, we may compare by measured parameters. The *profile* is set of measured numbers for an architecture.

Definition 4.2

The **grounding profile** of a GA at (k, t, U, P) is

$$\text{GP}(\text{GA}) := \{\varepsilon_{\text{pres}}^{k,t}, \varepsilon_{\text{faith}}^{k,t}, \delta_{\text{faith}}^{k,t}, \omega_U^{k,t}, \delta_{\text{comp}}^{k,t}, \beta^{k,t}\}.$$

We compare grounding architectures by their profiles at (k, t, U, P) ; mode labels are descriptive. Examples: *symbolic* modes fix meanings by explicit rules inside a calculus (strong composition, weak world causation unless rules are learned); *referential* modes are anchored by perception and action, with etiological warrant from covariation learned under task success; *vectorial* modes locate meanings in vector geometry, where distances/directions track regularities, and world contact arises only if trained on grounded signals (e.g., paired modalities); *relational* modes give meaning by position in typed networks of relations and inferences, excellent for abstraction and rule-based composition, with world contact only if learned from data.

4.1 Symbolic grounding

Symbolic grounding obtains when symbols derive meaning from explicit symbolic structures (definitions, axioms, rules). In the limiting case, the representation space collapses to the symbol space ($\mathcal{R} = \Sigma$), and meanings are fixed by a stipulated interpretation $\mathcal{I}_k^t : \Sigma \rightarrow \mathcal{M}_k^t$ (e.g., a dictionary entry defining *BACHELOR* as “unmarried man”, or a logical knowledge base with model-theoretic semantics [40]). This is intra-symbolic grounding: symbols explained by other symbols and formal relations.

Unless otherwise stated we take symbolic systems at

$$E = (k, t, U, P) = (k_{\text{logic}}, \inf, U_{\text{edit}}, P_{\text{strings}}),$$

where $\mathcal{R} = \Sigma$ with edit-distance $d_{\mathcal{R}}$, threat model U_{edit} comprises character-level perturbations scaled by edit distance (e.g., single-edit typos), and P_{strings} is a reference distribution over \mathcal{R} .

A canonical instantiation takes: Σ as symbols (words, predicates), $\mathcal{R} = \Sigma$ (trivial representation), \mathcal{C} as a conceptual-role/derivation space, $\mathcal{M}_k^{\text{ext}}$ as inferential/intensional meanings, $\Phi = \text{id}_{\Sigma}$, and $\Gamma : \Sigma \rightarrow \mathcal{C}$ as table-lookup/rule-closure, so $\Psi = \Gamma$.

Definition 4.3

A mode is **symbolic** at $(k, t) = (k_{\text{logic}}, \inf)$ if:

1. $(\mathcal{R}, d_{\mathcal{R}})$ is uniformly discrete; take $\mathcal{R} = \Sigma$ with edit metric;
2. there exists an analyst-given $\mathcal{I}_k^t : \Sigma \rightarrow \mathcal{M}_k^t$ such that, for all atoms $\sigma \in \Sigma_{\text{atom}}$,

$$\mathcal{A}_k^t \Psi(\sigma) = \mathcal{I}_k^t(\sigma)$$

(with $\Phi = \text{id}$ and Γ rule/lookup);

3. G0-weak holds; G0-strong holds only if Φ, Γ and any appealed \mathcal{A}_k^t were acquired under selection process \mathcal{T} .

For intensional/inferential evaluation, we take $\mathcal{A}_k^t : \mathcal{C} \rightarrow \mathcal{M}_k^t$ to identify conceptual roles with meanings (or to map them explicitly if types differ). By fiat, \mathcal{I}_k^t agrees with $\mathcal{A}_k^t \Psi$ on atoms. When the grammar supplies typed constructors and the semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ interprets each constructor f by $f_k^{\mathcal{M}, t}$, and when Γ realizes the grammar, the homomorphic extension satisfies (notated with the uparrow on \mathcal{I}_k^t)

$$\mathcal{A}_k^t \Psi(f(\vec{\sigma})) = f_k^{\mathcal{M}, t}(\mathcal{A}_k^t \Psi(\sigma_1), \dots, \mathcal{A}_k^t \Psi(\sigma_n)) = \mathcal{I}_k^{t, \uparrow}(f(\vec{\sigma})),$$

so G1 holds on atoms ($\varepsilon_{\text{pres}}^{k, t} = 0$) and G2a holds on all well-typed composites in-grammar ($\varepsilon_{\text{faith}}^{k, t} = 0$). Absent a documented selection process \mathcal{T} that selected for task success, symbolic systems with purely stipulated mappings fail G2b (no selected-for history), even if ablations show causal dependence.

For G3, since $(\mathcal{R}, d_{\mathcal{R}})$ is uniformly discrete with minimal nonzero distance 1, any map $S_{k, t} = \mathcal{A}_k^t \circ \Gamma$ is uniformly continuous with a degenerate global modulus: $\omega(\varepsilon) = 0$ for $\varepsilon < 1$ and $\omega(\varepsilon) \leq \text{diam}(S_{k, t}(\mathcal{R}))$ for $\varepsilon \geq 1$. Under U_{edit} (single-edit typos), we expect empirical curves $\hat{\omega}_U^{k, t}$ to jump from 0 to near-maximal semantic deviation at the first non-zero radius: *cat* to *car* can induce a *large* intensional shift!

For G4, compositionality is perfect *inside* the grammar ($\delta_{\text{comp}}^{k, t} = 0$). For out-of-grammar combinations (e.g., lacking a lexical entry or constructor for a modifier in something like *quantum bachelor*), systematicity on the held-out we expect collapse ($\beta^{k, t} \approx 0$), unless rules are added.

Symbolic grounding, as characterized here, satisfies G0-weak, G1, and G4 (on-grammar), but fails G2b without a selection history, and offers no practically useful G3 robustness to small input perturbations under U_{edit} . Meanings reside in the stipulated interpretation and calculus: a scope mismatch with grounding requirements rather than a criticism of formal semantics per se.

4.2 Referential grounding

Referential grounding connects symbols to entities and properties in the world through sensorimotor experience: meanings are anchored by causal interaction rather than by intra-symbolic definitions. The paradigm case is a child learning *cat* via encounters with actual cats; perceptual regularities and successful actions stabilize the category.

We evaluate at

$$E = (k, t, U, P) = (k_{\text{embodied}}, \text{ext}, U_{\text{nat}}, P_{\mathcal{R}}),$$

where U_{nat} consists of naturalistic sensory perturbations (e.g., lighting, pose, viewpoint, sensor noise) with a scale $\varepsilon(u)$, and $P_{\mathcal{R}}$ is a reference distribution over \mathcal{R} (e.g., in-distribution sensorimotor states). When considered, we report an adversarial subset U' (e.g., adversarial patches or carefully crafted distractors) separately.

Within our framework, a canonical instantiation: Σ as linguistic tokens/commands (e.g., “red ball”); \mathcal{R} as sensorimotor states/features (the extent to which these are possibly *fused* from language and perception remains open [34]); \mathcal{C} as perceptual/conceptual categories; and $\mathcal{M}_k^{\text{ext}}$ as task/world states (entities, properties, relations) for the declared context k . The encoder Φ provides the causal portal from surface symbols to internal representations (and, in interactive settings, \mathcal{R} also receives concurrent sensory input); the map $\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ induces task-level concepts; realized semantics for evaluation flows through

$$S_{k,t} = \mathcal{A}_k^{\text{ext}} \circ \Gamma \quad \text{and} \quad \Psi = \Gamma \circ \Phi \text{ for the symbol-conditioned pathway.}$$

The alignment $\mathcal{A}_k^{\text{ext}} : \mathcal{C} \rightarrow \mathcal{M}_k^{\text{ext}}$ ties internal concepts to world-level meanings and is expected, under G0-strong, to be an *agent-internal, learned* mechanism. The intended interpretation $\mathcal{I}_k^{\text{ext}}$ fixes task-relevant denotations in context k .

Definition 4.4

A mode is **referential** at $(k, t) = (k_{\text{embodied}}, \text{ext})$ if:

1. \mathcal{R} includes sensorimotor encodings relevant to k and $\mathcal{M}_k^{\text{ext}}$ denotes world/task states;
2. $\mathcal{A}_k^{\text{ext}} : \mathcal{C} \rightarrow \mathcal{M}_k^{\text{ext}}$ is an *internal* map whose parameters were *acquired under* an interactional/embodied process \mathcal{T} , establishing stable causal covariation between internal states and worldly regularities;
3. there exists an agent-internal mechanism M (within Φ, Γ , and/or $\mathcal{A}_k^{\text{ext}}$) acquired under \mathcal{T} such that, for a task success predicate $\text{succ} : \mathcal{O} \rightarrow \{0, 1\}$ and evaluation E ,

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t}.$$

In this sense, G2b (etiological faithfulness) holds for extensional tasks.

G0-strong holds when Φ, Γ (and any appealed $\mathcal{A}_k^{\text{ext}}$) are implemented inside the agent *and* were acquired by agent-internal learning or evolution \mathcal{T} that shaped perceptual-motor pathways; G0-weak holds if these mechanisms are implemented internally. An external measurement that supplies $\mathcal{A}_k^{\text{ext}}$ would violate G0-strong and must be declared G0-weak.

For concrete atoms $\sigma \in \Sigma_{\text{atom}}$ (e.g., basic object/color terms) with $\sigma \in \text{dom}(\mathcal{I}_k^{\text{ext}})$,

$$d_{k,t}(\mathcal{A}_k^{\text{ext}}\Psi(\sigma), \mathcal{I}_k^{\text{ext}}(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t}.$$

Intuitively, hearing red or mug should activate an internal concept that projects, via $\mathcal{A}_k^{\text{ext}}$, to the same region of world-states as the intended denotation in k . In referential settings with well-learned perceptual categories, we expect that $\varepsilon_{\text{pres}}^{k,t}$ is small for such concrete atoms, though we admit drift across contexts if $P_{\mathcal{R}}$ shifts (e.g., unusual lighting).

For (possibly composite) $\sigma \in \text{dom}(\mathcal{I}_k^{\text{ext}})$,

$$d_{k,t}(\mathcal{A}_k^{\text{ext}}\Psi(\sigma), \mathcal{I}_k^{\text{ext}}(\sigma)) \leq \varepsilon_{\text{faith}}^{k,t},$$

so realized meanings correlate with intended denotations in k . The etiological clause (G2b) supplies the causal warrant absent from purely stipulated systems: there exists an internal M (e.g., a learned visual-grounding or spatial-relation module) acquired under \mathcal{T} such that turning M off measurably reduces success on the task family ($\text{ACE}_E(M) \geq \eta^{k,t}$). Direct intervention/ablation or, when infeasible, by lower bounds via retrain-without- M , invariance across environments, or instrumental variables, with assumptions and uncertainty reported, probe for this.

With naturalistic U_{nat} and $P_{\mathcal{R}}$, we consider a modulus $\omega_U^{k,t}$ satisfying, for confidence level $1 - \alpha$,

$$\Pr_{r \sim P_{\mathcal{R}}} \left[\sup_{\varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^{\text{ext}}\Gamma(r), \mathcal{A}_k^{\text{ext}}\Gamma(u \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

where “ $u \cdot r$ ” denotes applying the perturbation u in \mathcal{R} . For benign, naturalistic U_{nat} , grounded systems should show graceful degradation (small $\omega_U^{k,t}$ over a useful radius). In contrast, adversarial families U' induce large semantic shifts

at tiny input scales; we expect that if global Lipschitz fits fail, robustness should be reported locally or distributionally and adversarially *separately*.

Given typed constructors in the grammar and a semantic algebra $\langle \mathcal{M}_k^{\text{ext}}, \mathcal{F}_k^{\text{ext}} \rangle$, we assess

$$d_{k,t} \left(\mathcal{A}_k^{\text{ext}} \Psi(f(\vec{\sigma})), f_k^{\mathcal{M}, \text{ext}}(\mathcal{A}_k^{\text{ext}} \Psi(\sigma_1), \dots, \mathcal{A}_k^{\text{ext}} \Psi(\sigma_n)) \right) \leq \delta_{\text{comp}}^{k,t}.$$

Novel attribute-noun or relational combinations (e.g., *metallic balloon*, or spatial compositions unseen in training) elevate $\delta_{\text{comp}}^{k,t}$ and reduce systematicity $\beta^{k,t}$ unless prior exposure or structured inductive bias supports composition. Perceptual fusion (e.g., combining vision and haptics) may mitigate this by constraining the concept manifold, but it does not guarantee full compositionality [59].

Referential grounding excels on concrete, manipulable categories, supports cross-modal validation (e.g., visual and haptic agreement), and delivers etiological warrant in spades by tying content to histories of successful action. It is, however, limited for abstractions (*justice*, *democracy*), theoretical entities, or fictional kinds (e.g., *dragons*), which lack stable extensional anchors in $\mathcal{M}_k^{\text{ext}}$ and are better treated in inferential or social-normative modes [37]. The very requirement that content be selected-for by worldly success also marks its boundary for non-observables and culturally variable categories.

4.3 Vectorial grounding

Vectorial grounding maps symbols⁶ to points in high-dimensional spaces where geometric relations encode semantics [48]. In \mathfrak{G} , the representation space is $\mathcal{R} = \mathbb{R}^n$; $\Phi : \Sigma \rightarrow \mathbb{R}^n$ embeds tokens, $\Gamma : \mathbb{R}^n \rightarrow \mathcal{C}$ decodes to conceptual features, and $\Psi = \Gamma \circ \Phi$. For $(k_{\text{ling}}, \text{ext})$, $\mathcal{M}_k^{\text{ext}}$ comprises inferential/role-like targets (e.g., graded similarity, definitional features), and $\mathcal{A}_k^{\text{ext}} : \mathcal{C} \rightarrow \mathcal{M}_k^{\text{ext}}$ aligns concepts to those targets (often instantiated via evaluation protocols).

We evaluate at

$$E_{\text{ling}} = (k_{\text{ling}}, \text{ext}, U_{\text{ling}}, P_{\mathcal{R}}^{\text{ling}}) \quad \text{and, where noted,} \quad E_{\text{vl}} = (k_{\text{vl}}, \text{ext}, U_{\text{vl}}, P_{\mathcal{R}}^{\text{vl}}).$$

For robustness (G3) in the linguistic setting, U_{ling} comprises text perturbations such as typos, character/word-level noise, paraphrase templates; in multimodal, U_{vl} naturalistic image perturbations (crop, blur, color jitter) paired with benign text edits. We specify a reference distribution $P_{\mathcal{R}}$ over \mathcal{R} in each case (e.g., in-distribution corpora or image-text pairs).

A signature property is approximate *geometric alignment*:

$$d_{\mathcal{R}}(\Phi(\sigma_1), \Phi(\sigma_2)) \approx d_{k,t}(\mathcal{I}_k^t(\sigma_1), \mathcal{I}_k^t(\sigma_2)),$$

so relative positions in \mathbb{R}^n predict graded semantic proximity in the declared meaning space.

Definition 4.5

A mode is **vectorial** at $(k, t) \in \{(k_{\text{ling}}, \text{ext}), (k_{\text{vl}}, \text{ext})\}$ if:

1. $\mathcal{R} = \mathbb{R}^n$ with a norm metric $d_{\mathcal{R}}$, and $\Phi : \Sigma \rightarrow \mathbb{R}^n$ is learned under a process \mathcal{T} relevant to (k, t) ;
2. there exist an evaluation alignment \mathcal{A}_k^t and a tolerance $\varepsilon_{\text{geom}}^{k,t}$ such that, on a declared pair distribution,

$$|d_{\mathcal{R}}(\Phi(\sigma_i), \Phi(\sigma_j)) - d_{k,t}(\mathcal{I}_k^t(\sigma_i), \mathcal{I}_k^t(\sigma_j))| \leq \varepsilon_{\text{geom}}^{k,t};$$

3. G0-strong holds for $(k_{\text{ling}}, \text{ext})$ when Φ, Γ (and any appealed $\mathcal{A}_k^{\text{ext}}$) are acquired by self-supervised linguistic training under \mathcal{T} ; for (ext) claims, paired/interactive signals tied to the world are required for G0-strong with respect to the evaluated tasks.

For $(k_{\text{ling}}, \text{ext})$, self-supervised training (e.g., distributional objectives, masked/next-token modeling) implies a G0-strong insofar as they are acquired internally under \mathcal{T} and evaluated on linguistic inference targets. For $(k_{\text{vl}}, \text{ext})$, world authenticity requires signals tied to the world (interaction, grounded supervision, paired modalities); otherwise, only G0-weak holds for claims. If \mathcal{A}_k^t is supplied externally, any sort of extensional grounding claims must be labeled G0-weak unless itself is learned internally under \mathcal{T} .

For atomic items $\sigma \in \Sigma_{\text{atom}}$ at $(k_{\text{ling}}, \text{ext})$,

$$d_{k,t}(\mathcal{A}_k^{\text{ext}} \Psi(\sigma), \mathcal{I}_k^{\text{ext}}(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t},$$

⁶For $(k_{\text{vl}}, \text{ext})$, co-embedding models (e.g., CLIP) align visual and textual vectors to shared structure [58].

with $\varepsilon_{\text{pres}}^{k,t}$ estimated on a declared test set. Beyond atoms, vector geometry tracks human-like similarity and definitional features to a useful degree, for which we expect yielding small $\varepsilon_{\text{faith}}^{k,t}$; note that fine-tuning warps or sharpens these axes [60]. Selection history \mathcal{T} for standard language models optimizes linguistic objectives (next-token likelihood, contrastive text-image alignment), supplying warrant particularly for linguistic correlational tasks; it does *not*, by itself, establish embodied/task success in (ext) unless training includes interaction or grounded feedback.

G2b requires an internal learned mechanism M (within Φ, Γ , and/or \mathcal{A}_k^t) whose contribution to the evaluated task family is causal and selected-for. We therefore report

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t},$$

with respect to a binary success predicate appropriate to E (e.g., correct pairwise ranking or classification for $(k_{\text{ling}}, \text{ext})$; correct retrieval/grounding for $(k_{\text{vl}}, \text{ext})$). Classic analogy patterns (e.g., the infamous $\Phi(\text{king}) - \Phi(\text{man}) + \Phi(\text{woman}) \approx \Phi(\text{queen})$) are correlational evidence (G2a); they do *not* establish G2b without intervention/ablation linking M to improved success under \mathcal{T} .

Distributed representations tend to yield smooth local semantics [19, 70]. Under U_{ling} (typos/paraphrase) or benign U_{vl} (small crops/blur), we situate a modulus $\omega_U^{k,t}$ with small slope near 0, a graceful degradation [15]. Note, however, that adversarial triggers or off-manifold edits induce large semantic shifts at tiny input scales [55, 19]; global Lipschitz bounds should fail, hence we encourage a separate report, in which robustness reported locally (or distributionally), and adversarial regimes separately. Formally, for confidence level $1 - \alpha$,

$$\Pr_{r \sim P_{\mathcal{R}}} \left[\sup_{\varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^t \Gamma(r), \mathcal{A}_k^t \Gamma(u \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

where “ $u \cdot r$ ” denotes applying the perturbation in \mathcal{R} .

Compositional behavior seems to depend on architecture. Static embeddings [46, 53] support limited additive composition. Contextual models combine token contributions via attention, yielding representations that integrate multiple properties and sometimes approximate compositional semantics [82, 22, 83, 24, 73]. We therefore assess, for typed constructors f and context k ,

$$d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), f_k^{\mathcal{M},t}(\mathcal{A}_k^t \Psi(\sigma_1), \dots, \mathcal{A}_k^t \Psi(\sigma_n))) \leq \delta_{\text{comp}}^{k,t},$$

and report systematicity $\beta^{k,t}$ on held-out compositional splits (e.g., unseen attribute-noun pairings, relational templates). LLMs, then, seem to achieve workable, albeit imperfect, $\delta_{\text{comp}}^{k,t}$ and non-trivial $\beta^{k,t}$; failures concentrate on out-of-distribution combinations and deep nesting.

Vectorial grounding’s glory is G2a (correlational fit) and local G3 (smoothness), which should not be surprising, given that the very DNA of vectorial models is just that: vectors with geometric behavior, and supports similarity judgments (*cat* \approx *tiger* \neq *truck*), and enables few-shot generalization via geometric regularities. Its costs include opacity (hard-to-interpret Γ), compositional fragility (attention mechanisms approximate but do not guarantee algebraic composition [24]), and limited denotational capacity: vectors encode statistical correlates of referents but, without world-anchored learning, do not *denote* particular entities or secure G2b for embodied tasks. Multimodal contrastive training (e.g., CLIP) partially closes this by blending vectorial and referential signals [58], improving G2a in (ext) settings and, with appropriate supervision, moving toward (but not guaranteeing) G2b.

4.4 Relational grounding

Relational grounding establishes meaning through conceptual roles and inference patterns. Unlike vectorial grounding, which uses geometry in \mathbb{R}^n , relational grounding uses discrete structure (types, roles, subsumptions, constraints), and aims at Fregean sense, the mode of presentation, supporting meaning for abstraction and fictional entities that lack extensional denotation (e.g., Frege’s “least rapidly convergent series”, or *dragon*).

We evaluate at

$$E = (k, t, U, P) = (k_{\text{logic}}, \text{ext}, U_{\text{rel}}, P_{\mathcal{R}}),$$

where representations are discrete, graph-, or rule-structured: \mathcal{R} is a set of nodes/edges/axioms with a graph/edit metric $d_{\mathcal{R}}$. For robustness (G3) we declare U_{rel} consisting of local symbolic edits (edge swaps, axiom insertion/deletion, property retyping) with some scale $\varepsilon(u)$ given by the number/type of edits, and specify a reference distribution $P_{\mathcal{R}}$ over \mathcal{R} (e.g., in-distribution subgraphs/ontologies).

We instantiate: Σ as symbols (e.g., *bachelor*, *dragon*); \mathcal{R} as concept nodes/edges and axioms; \mathcal{C} as a role space; and $\mathcal{M}_k^{\text{ext}}$ as inferential/intensional meanings. The encoder $\Phi : \Sigma \rightarrow \mathcal{R}$ maps symbols to conceptual nodes/schemata;

$\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ computes rule closure (entailments/constraints); thus $\Psi = \Gamma \circ \Phi$ is realized conceptual content. For evaluation, we set $\mathcal{A}_k^{\text{ext}} : \mathcal{C} \rightarrow \mathcal{M}_k^{\text{ext}}$ to an identification of role content with meanings (typed if needed). Meaning emerges from inferential patterns: for $r = \Phi(\text{“bachelor”})$, $\Gamma(r)$ yields $\{\text{HUMAN, MALE, } \neg\text{MARRIED}\}$; for taxonomy links $\text{cat} \sqsubseteq \text{animal}$, composition is rule-based.

Definition 4.6

A mode is **relational** at $(k, t) = (k_{\text{logic}}, \text{ext})$ if:

1. \mathcal{R} is a typed relational/graphical structure (e.g., a first-order structure or description-logic KB) with edit/graph metric $d_{\mathcal{R}}$;
2. $\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ computes intensional role sets via closure, and $\mathcal{A}_k^{\text{ext}}$ identifies role content with $\mathcal{M}_k^{\text{ext}}$;
3. there exists an internal mechanism M (within Φ, Γ , and/or $\mathcal{A}_k^{\text{ext}}$) acquired under a process \mathcal{T} that selects for inferential success on the declared task family, such that the average causal effect

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t},$$

where succ is a logical/inferential success predicate (e.g., correct entailment or QA over the ontology). No claim is made about (ext) tasks.

G0-weak holds for hand-engineered ontologies (internal mechanisms without acquisition). G0-strong holds when Φ, Γ (and any appealed $\mathcal{A}_k^{\text{ext}}$) are acquired by an internal process \mathcal{T} (e.g., relation extraction/induction from text or data). In practice, curated KGs meet only G0-weak; we would expect that a learned neuro-symbolic (see [66, 8]) meets G0-strong.

When the intended interpretation $\mathcal{I}_k^{\text{ext}}$ is specified in the same role language, tolerances can be tight. For atoms $\sigma \in \Sigma_{\text{atom}} \cap \text{dom}(\mathcal{I}_k^{\text{ext}})$,

$$d_{k,t}(\mathcal{A}_k^{\text{ext}}\Psi(\sigma), \mathcal{I}_k^{\text{ext}}(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t}.$$

For defined composites,

$$d_{k,t}(\mathcal{A}_k^{\text{ext}}\Psi(\sigma), \mathcal{I}_k^{\text{ext}}(\sigma)) \leq \varepsilon_{\text{faith}}^{k,t}.$$

Intuitively, consider: if the ontology is complete for the tested fragment, then the agent’s computed role content should coincide with the stipulated roles (up to the declared tolerance); gaps or inconsistencies in the relation set show up as preservation or faithfulness errors.

For inferential tasks (see [3, 79]), G2b holds only when rules/relations were selected for under \mathcal{T} , because they improve success (coverage/precision) on the task family, and when a learned internal M exhibits $\text{ACE}_E(M) \geq \eta^{k,t}$. Hand-crafted ontologies rarely document such selection histories [43, 17]; learned extractors can satisfy G2b for *inferential* tasks. By design, no claim is made about embodied/world success in \mathcal{M}^{ext} .

With uniformly discrete \mathcal{R} , uniform continuity is automatic (Corollary A.1), but the global modulus can be degenerate. Formally, for confidence level $1 - \alpha$,

$$\Pr_{r \sim P_{\mathcal{R}}} \left[\sup_{u \in U_{\text{rel}} : \varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^{\text{ext}}\Gamma(r), \mathcal{A}_k^{\text{ext}}\Gamma(u \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

and “ $u \cdot r$ ” applying a symbolic edit u to r . Dense, type-constrained neighborhoods imply small local moduli (that is, graceful degradation); sparse or brittle regions imply large jumps where tiny edits cause major semantic shifts (e.g., removing a single subsumption edge breaks multiple entailments [69]). Robustness should therefore be reported locally and with explicit U_{rel} .

Relational systems excel at typed composition. Let the grammar provide constructors (e.g., typed conjunction, role restriction) and $\langle \mathcal{M}_k^{\text{ext}}, \mathcal{F}_k^{\text{ext}} \rangle$ interpret them. Then

$$\delta_{\text{comp}}^{k,t} = 0 : \quad \mathcal{A}_k^{\text{ext}}\Psi(f(\vec{\sigma})) = f_k^{\mathcal{M}, \text{ext}}(\mathcal{A}_k^{\text{ext}}\Psi(\sigma_1), \dots, \mathcal{A}_k^{\text{ext}}\Psi(\sigma_n)),$$

and systematicity $\beta^{k,t} \approx 1$ holds for held-out in-grammar combinations when rules apply uniformly. For example,

$$\Gamma \circ \Phi(\text{“red dragon”}) = \text{COLOR : RED} \sqcap \text{TYPE : DRAGON}.$$

Failures follow from incompleteness or type violations rather than from the algebra itself.

Relational modes have liabilities: dependence on completeness/consistency of the relation set (missing axioms break chains); brittleness under small symbolic edits (G3 outside dense neighborhoods); lack of extensional denotation, so G2b does not target embodied/world success. Compared to vectorial grounding, relational is explicit but less tolerant of sparsity; it trades causal world contact for inferential clarity. This is a non-negligible motivation for neurosymbolic systems, which blend relational structure with learned components (e.g., learned extractors for G0-strong and G2b on inferential tasks). Vector-referential models such as CLIP [58] lie outside purely relational methods, but illustrate how adding signals can improve G1–G3 in (ext) evaluations, complementing the intensional strengths of relational grounding.

5 How is model-theoretic semantics grounded?

We evaluate *model-theoretic semantics* (MTS) (of [85, 25]) at $(k, t) = (k_{\text{logic}}, \text{inf})$. The representation space is uniformly discrete: $\mathcal{R} = \Sigma$ (syntactic items, types, well-formed formulae) with an edit/structural metric $d_{\mathcal{R}}$ (e.g., token-edit distance; minimum nonzero distance = 1). For robustness (G3), a threat model we are interested in U_{edit} (single-edit typos, token substitutions, minor parse edits) and a reference distribution $P_{\mathcal{R}}$ over well-formed Σ . The evaluation tuple is

$$E = (k_{\text{logic}}, \text{inf}, U_{\text{edit}}, P_{\mathcal{R}}).$$

MTS fixes truth conditions by postulating an analyst-given interpretation $\mathcal{I}_k^{\text{inf}} : \Sigma \rightarrow \mathcal{M}_k^{\text{inf}}$ that maps syntactic items to denotations in a mathematical structure: if $\Sigma = \{\text{cat}\}$, a model contains an element c with $\mathcal{I}_k^{\text{inf}}(\text{cat}) = c$. This is exactly ideal for logical inference, but it does not explain *why* that assignment should bind an agent to the world, and, hence, it is “a theory of *inference*, not of *meaning*” [54].

Formally, when the internal pipeline is trivial ($\Phi = \text{id}_{\Sigma}$, $\Gamma = \text{id}_{\Sigma}$, hence $\Psi = \Gamma \circ \Phi = \text{id}_{\Sigma}$ with $\mathcal{C} = \Sigma$) and $\mathcal{A}_k^{\text{inf}}$ identifies \mathcal{C} with $\mathcal{M}_k^{\text{inf}}$, we have, for atoms $\sigma \in \Sigma_{\text{atom}} \cap \text{dom}(\mathcal{I}_k^{\text{inf}})$,

$$d_{k,t}(\mathcal{A}_k^{\text{inf}}\Psi(\sigma), \mathcal{I}_k^{\text{inf}}(\sigma)) = 0,$$

so G1 holds with $\varepsilon_{\text{pres}}^{k,t} = 0$ by construction. The same identity yields trivial correlational faithfulness (G2a) wherever $\mathcal{I}_k^{\text{inf}}$ is total on the tested fragment; under a typed grammar and semantic algebra $\langle \mathcal{M}_k^{\text{inf}}, \mathcal{F}_k^{\text{inf}} \rangle$, the homomorphic extension $\mathcal{I}_k^{\text{inf},\uparrow}$ matches $\mathcal{A}_k^{\text{inf}}\Psi$ on all well-typed composites.

Because the key mappings are stipulated by that external analyst, MTS satisfies *only* G0-weak (mechanisms are implemented inside the calculus) and *not* G0-strong (no agent-internal acquisition under \mathcal{T}): nothing in the agents causal fabric earns the mapping. The etiological clause G2b requires a learned internal mechanism M (within Φ, Γ , and/or \mathcal{A}_k^t) with positive average causal effect

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t}.$$

MTS does not provide such selection-for-histories for worldly tasks, and thus fails G2b.

With uniformly discrete $(\mathcal{R}, d_{\mathcal{R}})$, $S_{k,t} = \mathcal{A}_k^t \circ \Gamma$ is uniformly continuous (Corollary A.1), but the global modulus is *degenerate*: $\omega(\varepsilon) = 0$ for $\varepsilon < 1$ and $\omega(\varepsilon) \leq \text{diam}(S_{k,t}(\mathcal{R}))$ for $\varepsilon \geq 1$. Under U_{edit} , $\hat{\omega}_{U_{\text{edit}}}^{k,t}$ therefore jumps at the first nonzero radius: a single-edit perturbation induces an arbitrary semantic change (ill-formedness, different symbol), so there is no useful robustness radius. The map is continuous only in the formal sense (Theorem A.1), but not graceful.

Perhaps unsurprisingly, given a typed grammar and semantic algebra $\langle \mathcal{M}_k^{\text{inf}}, \mathcal{F}_k^{\text{inf}} \rangle$, MTS satisfies exact homomorphic compositionality (Theorem A.2) as Frege and, later, Montague, intended: for all well-typed constructors f ,

$$\mathcal{A}_k^{\text{inf}}\Psi(f(\vec{\sigma})) = f_k^{\mathcal{M},\text{inf}}(\mathcal{A}_k^{\text{inf}}\Psi(\sigma_1), \dots, \mathcal{A}_k^{\text{inf}}\Psi(\sigma_n)),$$

with $\delta_{\text{comp}}^{k,t} = 0$ and $\beta^{k,t} = 1$ on the grammars closure. Compositionality preserves the stipulated atoms; it does not, by itself, however, ground them. This is framed by [28] as a “*second-order* symbol-grounding problem”: even if a logician ties *cat* to a model element c , what ties c to *actual* cats? MTS thus remains a descriptive calculus for analysts.

6 How are large language models grounded?

LLMs are not monolithic. Encoder-based *embedding models* (e.g., sentence transformers) are trained with contrastive objectives and produce pooled, normalized sentence vectors; *multimodal dual encoders* (e.g., CLIP) co-embed text and images; *decoder-only* models (e.g., OPT, GPT) are autoregressive and optimize next-token prediction rather than pooled

sentence embeddings; *encoder-only* models (e.g., BERT/RoBERTa) use masked-LM pretraining without contrastive objectives. Deliberately misquoting Saussure: in (large) language (models), there is only difference.

The debate over LLM grounding is active [36, 31, 52, 35]. [6] argue that text-only training precludes “real understanding”, while others claim pragmatic grounding can suffice for linguistic competence [28]. Within our framework, LLMs realize *vectorial* grounding for $(k_{\text{ling}}, \text{inf})$, and multimodal variants add partial *extensional* signals.

Unless noted, we evaluate at

$$E_{\text{ling}} = (k_{\text{ling}}, \text{inf}, U_{\text{ling}} \cup U_{\text{prompt}}, P_{\text{text}})$$

with linguistic meaning and text perturbations. U_{ling} comprises typos, token/word-level noise, and paraphrase templates; U_{prompt} comprises prompt perturbations and jailbreak-style adversarials. For multimodal LLMs, we are additionally armed

$$E_{\text{vl}} = (k_{\text{vl}}, \text{ext}, U_{\text{vl}}, P_{\text{img-text}})$$

with benign image transforms (crop/blur/jitter) paired with minor text edits.

Φ (embedding/contextualization) and Γ (decoding to conceptual features) are implemented in network weights; mechanisms are internal, so G0-strong holds for linguistic mechanisms under a self-supervised training \mathcal{T} . For extensional claims, G0-strong requires world-anchored signals (interaction, grounded labels, paired modalities); otherwise, only G0-weak holds for (ext) claims.

Many tokens map to paraphrase-stable meanings on average⁷. At $(k, t) = (k_{\text{ling}}, \text{inf})$ and a test distribution Q ,

$$\Pr_{\sigma \sim Q} \left[d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t} \right] \geq 1 - \eta.$$

Embedding geometry broadly tracks human similarity judgments [1, 64]:

$$d_{\mathcal{R}}(\Phi(\text{cat}), \Phi(\text{tiger})) \approx d_{k,t}(\mathcal{I}_k^t(\text{cat}), \mathcal{I}_k^t(\text{tiger})),$$

supporting small $\varepsilon_{\text{faith}}^{k,t}$ on similarity/definition protocols when in-distribution.

Selection history \mathcal{T} optimizes linguistic objectives (next-token likelihood, masked modeling, contrastive alignment), so we extrapolate: for linguistic proxy success predicates (e.g., dev loss, entailment/analogy accuracy), G2b holds, provided an internal mechanism M exhibits positive causal effect:

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t}.$$

RLHF/instruction-tuning aim to reduce $\varepsilon_{\text{faith}}^{k,t}$ on dialogue-style tasks; links to truth conditions or embodied control remain indirect unless interaction/grounded feedback is part of \mathcal{T} . For $(k_{\text{vl}}, \text{ext})$, multimodal contrastive training supplies partial world-anchored signals (improving G2a) but does not guarantee G2b without interaction.

Transformer representations exhibit local smoothness under token noise/orthographic variants [78, 49]. Adversarial prompts or off-manifold edits produce large semantic shifts at tiny input scales; hence, no global Lipschitz bound should be assumed (Theorem A.1). The key consequence of this: robustness should be reported *locally*, per threat.

LLMs achieve targeted compositional productivity [82], yet systematic generalization lags symbolic baselines on out-of-scope constructions [61]. We therefore measure, for typed constructors f and context k ,

$$d_{k,t}(\mathcal{A}_k^t \Psi(f(\vec{\sigma})), f_k^{\mathcal{M},t}(\mathcal{A}_k^t \Psi(\sigma_1), \dots, \mathcal{A}_k^t \Psi(\sigma_n))) \leq \delta_{\text{comp}}^{k,t},$$

and report $\beta^{k,t}$ on strict, held-out compositional splits. Compositional behavior is architecture-dependent (contextual attention versus static embeddings) and task-sensitive; ceiling effects in our current fragment (several models at $\beta \approx 1.0$) motivate stricter splits.

Succinctly: direct referential grounding is partial at best for text-only models; it emerges indirectly via text-world regularities or via added modalities (e.g., CLIP-style co-embedding [58]). Within our graded framework, LLMs occupy a kind of middle region: stronger than purely symbolic systems (G1/G2a/G3 local), but weaker than fully embodied agents on G2b for world tasks. They are improvable in the following: (i) adding interaction/embodiment to supply etiological warrant for \mathcal{M}^{ext} ; (ii) strengthening typed compositional generalization via explicit inductive biases and evaluation on strict splits (a dream of neurosymbolic models). In this way, we reconcile critiques that otherwise emphasize the absence of direct denotation [6, 31] with views stressing pragmatic grounding for linguistic competence [28], and provision metrics for measuring progress.

⁷In this case, we therefore report quantiles rather than maxima to account for hallucination/long-tail drift.

7 How is natural language grounded?

Consider a cat. What one person “knows” is a cat is not necessarily the same (by whichever metric we might check) as what another person “knows” as a cat; and yet, each interlocutor can have meaningful understanding about cats, whatever that may be. Grounding is influenced by anthropocentrism, and human language is the benchmark for grounding, though the *locus* of meaning remains contested. Correspondence-leaning views emphasize mapping words onto extra-linguistic reality [6]; pragmatic/inferentialist views stress that meaning arises from use within social practice [9, 10]. Our framework accommodates by typing meanings as extensional, inferential, or social-normative, and by indexing all claims to context k . We do not present measurements for human language grounding here; instead, we apply our grounding principles *in general* to predict behavior.

We frame human language in three settings:

$$\begin{aligned} E_{\text{ext}} &= (k_{\text{human}}, \text{ext}, U_{\text{ext}} \cup U_{\text{ling}}, P_{\text{human}}), \\ E_{\text{inf}} &= (k_{\text{human}}, \text{inf}, U_{\text{ling}}, P_{\text{human}}), \\ E_{\text{soc}} &= (k_{\text{human}}, \text{soc}, U_{\text{soc}} \cup U_{\text{ling}}, P_{\text{human}}), \end{aligned}$$

where U_{ling} includes accent variation, channel noise, and typographical slips; U_{ext} includes lighting/occlusion/viewpoint that affect perception and reference; and U_{soc} includes pragmatic mismatches (e.g., irony cues or register shifts). P_{human} denotes ordinary communicative situations (speaker, time, place, discourse state).

The causal machinery linking speech and text to perception, action, memory, and social practice is, not least of all, internally realized (neural/motor pathways) and acquired through evolutionary and developmental processes. Hence, G0-strong holds: Φ, Γ and the alignments \mathcal{A}_k^t are implemented inside the agent and acquired under \mathcal{T} (learning histories).

With context supplied, atomic expressions preserve intended meanings, within tolerances. For any $(k, t) \in \{(k_{\text{human}}, \text{ext}), (k_{\text{human}}, \text{inf}), (k_{\text{human}}, \text{soc})\}$ and $\sigma \in \Sigma_{\text{atom}} \cap \text{dom}(\mathcal{I}_k^t)$,

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{pres}}^{k,t}.$$

Polysemy and ambiguity enlarge these tolerances when k underspecifies the intended sense; discourse typically restores small errors via disambiguation.

For composed expressions (phrases, clauses) with $\sigma \in \text{dom}(\mathcal{I}_k^t)$,

$$d_{k,t}(\mathcal{A}_k^t \Psi(\sigma), \mathcal{I}_k^t(\sigma)) \leq \varepsilon_{\text{faith}}^{k,t},$$

reflecting that realized meanings track intended meanings under the given context and type.

Content is *selected for* through developmental learning and social feedback. Children who mismap categories (e.g., calling cats “dog”) experience correction and environmental counter-evidence; in adulthood, social norms regulate usage (e.g., sanctioning misuse of “liable” versus “likely”) [74, 4]. Formally, for outcomes \mathcal{O} appropriate to (k, t) (e.g., correct reference, successful joint action, normappropriate contribution) and success succ , there exists a learned internal mechanism M (within Φ, Γ , and/or \mathcal{A}_k^t) such that

$$\text{ACE}_E(M) := \mathbb{E} [\Pr(\text{succ} \mid \text{do}(M=\text{on})) - \Pr(\text{succ} \mid \text{do}(M=\text{off}))] \geq \eta^{k,t},$$

with the expectation taken over P_{human} (and, if relevant, over developmental processes \mathcal{T}). In ordinary settings, we predict positive causal contributions for mechanisms that support reference, inference, and pragmatic coordination.

On benign perturbations, human communication degrades gracefully. For confidence level $1 - \alpha$ and each declared U ,

$$\Pr_{r \sim P_{\text{human}}} \left[\sup_{u \in U: \varepsilon(u) \leq \varepsilon} d_{k,t}(\mathcal{A}_k^t \Gamma(r), \mathcal{A}_k^t \Gamma(u \cdot r)) \leq \omega_U^{k,t}(\varepsilon) \right] \geq 1 - \alpha,$$

with small slope near 0 (Theorem A.1): stability of meaning under small changes. Catastrophic slips (e.g., “desert” versus “dessert”) and perceptual illusions show that no global bound holds; robustness is local to high-probability neighborhoods of P_{human} and depends on context k and practice. Multimodal redundancy and pragmatic repair further sustain robustness [80, 81].

Speakers routinely interpret novel compounds (“quantum fridge”, “purple Tuesday”), nested clauses, and productive morphology, yielding low $\delta_{\text{comp}}^{k,t}$ and high systematicity $\beta^{k,t}$. Departures (idioms, coercion, adjectiveordering preferences) signal specialized constructions rather than wholesale failure of composition; typed and context-indexed constructors capture much of this productivity within our measurement scheme.

Items like “justice”, “quark”, or “Sherlock Holmes” lack direct perceptual anchors; their grounding is primarily inferential or social-normative, not extensional. Cultural and lexical variation imply that $d_{k,t}$ *cannot* be universal; for instance, what counts as a “river” in English may split into distinct lexicalizations in Yélf Dnye [41, 56]. Such facts enlarge tolerances in the relevant (k, t) , but do not break the grounding structure.

Across $(k_{\text{human}}, \text{ext/inf/soc})$ and everyday threat models, human linguistic agents communicate, and tend to know what they are talking about. In this way, agents achieve tight tolerances on G1–G4 under G0strong: authenticity is biological and learned; faithfulness is shaped by developmental and social selection; robustness is sustained by multimodal redundancy and pragmatic repair; compositionality is built into grammar and usage. Theories differ on *where* semantic links reside (world, practice, or both), but converge on the empirical point that, for most daily communication, $\mathcal{A}_k^t \Psi(\sigma) \approx \mathcal{I}_k^t(\sigma)$ within small, context-indexed tolerances. Natural language is, therefore, the paradigmatic grounded system, and remains that against which artificial architectures are compared, be that a fair comparison or otherwise.

8 Discussion, limitations, and future work

Our aim here for this framework is to admit evaluation on both conceptual fidelity and empirical utility. Our work here is intentionally idealized, and we invite further work and cooperation in experimental application.

8.1 Drawing hands

Our symbolic starting point may itself be questioned by embodied and enactive approaches to cognition. From an enactivist perspective, meaning *emerges* from organism-environment interactions, and only secondarily do we see symbols [77, 20]. An important limitation follows: our modes remain formal precisely because they treat symbols as the basic currency; we are presuming, then, a vantage of symbols at all.

Where we ask “how do symbols Σ ground in meanings \mathcal{M} ?”, enactivism asks the prior question: “how do stable symbolic distinctions emerge from fluid sensorimotor dynamics?” We begin where theirs ends: already-crystallized symbols awaiting semantic attachment. In this sense, we may be analyzing the *result* of grounding rather than grounding itself; a crystallographer studies ice while agnostic about the phase transition from water.

This clarifies our contribution’s scope. Most deployed AI systems, from theorem provers to large language models, *do* operate on discrete symbols or tokens, regardless of their ultimate metaphysical status [16, 50, 51, 14]; the question of how symbols relate to meanings remains pressing, even if symbols are themselves emergent. We provide measurable criteria for this relationship without requiring commitment to symbols as ontologically primitive. The etiological warrant (G2b) even gestures toward enactivist concerns by requiring that semantic mappings be earned through interaction, not stipulated.

A complete theory should (eventually) ground our framework itself: treating symbol emergence as a grounding problem where continuous dynamics (a new \mathcal{R}) ground discrete symbolic structures (a new \mathcal{M}), with tolerances capturing how much dynamics can vary while maintaining symbolic stability. Such an extension would nest our account within a broader enactive framework, rather than opposing it. Until then, we operate in the space between: symbols exist, however they arose, and their semantic coordination remains both problematic and measurable.

Recall *Drawing Hands* [23]. Our work here is Escher-like insofar as we use symbols to define symbols, a metalanguage to talk about an object language, proofs about mappings between the two. Left untended, we devolve into Harnad’s “symbol merry-go-round”, in which definitions chase definitions. Our way out is to keep the levels straight and require at least one downward arrow that is not itself symbolic; in our terms, G0 (authenticity) and G2b (selection-for) insist that some parts of the pipeline, be they perception, action, social uptake, live inside the agent and were shaped by success in a task family. That breaks the cycle: the metasymbols in this paper describe structures, but the meanings we evaluate are anchored by \mathcal{A}_k^t into non-symbolic contact (world, inference practice, norms). The mathematics then measures continuity, preservation, and composition of those mappings without needing to “borrow” semantics from the analyst at every step.

We move, therefore, from drawing hands to two hands connected to a body: the descriptive hand (our metalanguage) specifies profiles, moduli, and homomorphisms; the operative hand (the agent) grasps, predicates, repairs. When the only arrows are symbol \rightarrow symbol, then regress threatens; when some arrows run symbol \rightarrow representation \rightarrow concept \rightarrow world/practice and are earned by process, the loop closes pragmatically rather than paradoxically. We label what is stipulated (G0-weak), what is learned (G0-strong), and then audit preservation, faithfulness, robustness, and compositionality, and so spells out this break. Our study of symbols with symbols is a controlled vantage, and useful, so long as we keep one foot on the ground.

8.2 A typology of grounding profiles

The desiderata are independent, and systems can satisfy some while failing others; this independence generates a space of possible profiles which correspond to recognizable behavioral patterns. We sketch a typology here, both to illustrate the framework’s diagnostic power and to provide vocabulary for characterizing systems encountered in practice.

The interaction between correlational faithfulness (G2a) and compositionality (G4) yields a basic fourfold division is given in Table 1.

Table 1: Fourfold typology for correlational faithfulness and compositionality.

	High G4	Low G4
High G2a	<i>grounded</i>	<i>memorizer</i>
Low G2a	<i>miscalibrated</i>	<i>lost</i>

A system with high G2a and high G4 is *grounded* in the intuitive sense: it gets the right answers, and does so by combining meanings systematically. A system with high G2a but low G4 is a *memorizer*: it produces correct outputs for items it has seen, but its success depends on lookup rather than productive combination; such systems fail on novel compositions even when their constituents are familiar. Conversely, a system with low G2a but high G4 is *miscalibrated*: it applies the right algebra to wrong atoms, yielding systematic but incorrect results. Finally, a system low on both is simply *lost*: neither accurate nor systematic, it lacks the structure to ground meaning in any useful sense.

The distinction between correlational (G2a) and etiological (G2b) faithfulness separates systems that happen to be right from those that are right for the right reasons is given in Table 2.

Table 2: Fourfold typology for correlational faithfulness and etiological faithfulness.

	High G2b	Low G2b
High G2a	<i>competent</i>	<i>lucky</i>
Low G2a	<i>effortful failure</i>	<i>random</i>

A system satisfying both is *competent*: its accuracy is underwritten by internal mechanisms that were selected for success. A system with high G2a but low G2b is *lucky* or otherwise *shortcutting* byexploiting spurious correlations or surface heuristics that happen to align with correct answers on the evaluation distribution, but lack causal warrant; such systems are vulnerable to distribution shift precisely because their success was not earned. A system with high G2b but low G2a represents an *effortful failure*, in that it has learned mechanisms that demonstrably contribute to its behavior, but those mechanisms have latched onto the wrong targets; rescuing and debugging such a system requires retraining rather than recalibration. Low scores on both indicate a system operating essentially at *random* with respect to the task.

Robustness (G3) and compositionality (G4) distinguish architectural styles, given in Table 3.

Table 3: Fourfold typology for robustness and compositionality.

	High G4	Low G4
High G3	<i>smooth generalist</i>	<i>robust lookup</i>
Low G3	<i>brittle algebraist</i>	<i>fragile memorizer</i>

Systems high on both are *smooth generalists*, and combine meanings systematically and degrade gracefully under perturbation; this is the neural ideal for many applications. Systems with high G3 but low G4 are *robust lookup tables* that are stable under noise, but are otherwise inflexible, unable to extrapolate beyond stored patterns. Systems with high G4 but low G3 are *brittle algebraists*: they exhibit the perfect compositional behavior characteristic of symbolic systems (Theorem A.2), but the degenerate robustness modulus of discrete representations (Corollary A.1) makes them fragile under even minor perturbation, in which a single typo can be fatal or otherwise catastrophic, and induce arbitrary semantic change. Systems low on both are *fragile memorizers*, which, in the worst case, combine all the undesirable attributes of inflexibility with instability.

The interaction between authenticity (G0) and accuracy (G2a) separates genuine understanding from its simulacra, given in Table 4.

Table 4: Fourfold typology for robustness and compositionality.

	High G2a	Low G2a
G0-strong	<i>genuine</i>	<i>authentic failure</i>
G0-weak	<i>cargo cult</i>	<i>broken puppet</i>

A system satisfying G0-strong with high G2a exhibits what we might hazard to call (with appropriate delicacy) *genuine understanding*; a *genuine* system gets things right via mechanisms it has itself acquired through learning or evolution. A system with high G2a but only G0-weak is a *cargo cult*⁸: it produces correct outputs, but the semantic work is done by external stipulation rather than internal acquisition, and so the system behaves as if it understands, but meaning has been smuggled in by the designer or analyst⁹. A system satisfying G0-strong with low G2a is an *authentic failure*: it has genuinely learned something, but the wrong thing; this is not to say that genuine, authentic failures are not useful; rather, such failures are informative, as they reveal what the learning process actually selected for, a goal of this enterprise here. A system failing both is a *broken puppet*: neither accurate nor autonomous, it lacks both external and internal semantic grounding.

We may further combine the above to produce a cast of characters according to grounding that fulfill certain archetypes that recur across systems, shown in Table 5.

Table 5: Grounding archetypes, characterized by their respective typological behaviors.

Archetype	Profile	Example
<i>Parrot</i>	high G2a; low G2b/G4	lookup table; pattern-matching LLM
<i>Calculator</i>	high G4; low G1	perfect algebra on wrong atoms
<i>Glass canon</i>	high G2a; low G3	accurate; collapses under noise
<i>Fluent empty</i>	high G3/G4 (ling); low G2b (ext)	text-only LLM on world tasks
<i>Brittle expert</i>	high G2a/G2b; low G3	symbolic theorem prover
<i>Drifter</i>	low G1; cascading errors	corrupted embeddings
<i>Grounded</i>	high across all	human language (ideally)

The *parrot* reproduces correct outputs without causal warrant or systematic structure (a lookup table; an LLM exploiting surface patterns without genuine composition); in this way, we characterize the archetype’s namesake according to grounding [5]. The *calculator* applies perfect algebra to miscalibrated atoms, yielding systematic error. The *glass canon* achieves remarkable accuracy under standard conditions with no robustness margin. The *fluent empty* describes text-only language models: compositionally productive and locally smooth in the space of linguistic inference, yet lacking selection-for-success on world-referential tasks. The *brittle expert* characterizes symbolic theorem provers, in that such systems are correct and causally warranted within their domain, with no tolerance for noise. The *drifter* suffers corrupted atomic representations that propagate through composition, such that even perfect G4 cannot rescue a system whose G1 has failed. The *grounded* system, finally, scores high across all desiderata, which we take as its prototypical example to be natural language.

By locating a system within profile-space, as it were, we identify which desiderata require attention. A memorizer needs compositional structure; a miscalibrated system needs atomic recalibration; a lucky system needs causal validation; a brittle expert needs continuous relaxation or redundancy. The grounding profile thus functions as a diagnostic instrument, insofar as it tells us whether a system is grounded as well as how it fails to be, and, therefore, what interventions might help. Adequacy thresholds remain application-specific (see below), but the typology provides a common vocabulary for characterizing failures and prioritizing improvements.

Where a typology is useful is in what it typologizes; where a typology is interesting is where it breaks, or, rather, where what it attempts to typologize resists classification. A system that is simultaneously memorizer and miscalibrated, or that shifts cells under different (k, t) evaluations, reveals structure the typology did not anticipate. How does a memorizer become grounded? What interventions move a system from brittle algebraist to smooth generalist? The static typology invites a dynamics we do not develop here.

8.3 Metricization and calibration

The framework assumes well-defined (pseudo)metrics on all spaces: $d_{\mathcal{R}}$ on representations; $d_{k,t}$ on meanings; $d_{\mathcal{C}}$ on concepts. This assumption holds reasonably for extensional meanings grounded in vision (e.g., IoU-based metrics

⁸A term coined by Norris Mervyn Bird in 1945, in reference to the effects of colonial authorities in Melanesia.

⁹Recall: Searle’s Chinese Room: correct responses with no internal understanding.

for object reference) or for vectorial representations (cosine distance), but becomes non-trivial for social-normative meanings ($t = \text{soc}$). How should we metrize the space of dialogue commitments, normative statuses, or politeness violations? Candidate approaches include repair-cost metrics (the effort required to restore conversational alignment after a misfire), scalar judgments from human raters, or graph distances in networks of social obligations. Each choice encodes theoretical commitments about what makes meanings “close” or “distant” in social space, and no consensus exists. Similarly, for highly abstract or theoretical concepts, defining $d_{k,t}$ that respects intuitive similarity while remaining empirically tractable remains open. In practice, this may restrict the framework’s applicability to meaning types where reasonable metrics are established or can be operationalized through proxy tasks.

A related calibration problem concerns the tolerances $(\varepsilon_{\text{pres}}^{k,t}, \varepsilon_{\text{faith}}^{k,t}, \delta_{\text{comp}}^{k,t}, \text{etc.})$, which require empirical estimation without universal protocol. What counts as “small” preservation error depends on application: a medical diagnosis system may require very small $\varepsilon_{\text{pres}}^{k,t}$, for example, while a casual chatbot might tolerate comparatively larger values. We provide machinery for reporting these numbers, but limited guidance on what values constitute “good enough” grounding for particular tasks (indeed, far be it from us to state such a metric for the medical domain!). A more complete theory would connect grounding parameters to downstream task requirements systematically; we defer this to application-specific engineering judgments below.

8.4 Etiological variation

Etiological faithfulness (G2b) poses particular difficulty: verifying that internal states were “selected for” success requires access to training histories and causal interventions that may be infeasible for large-scale or proprietary systems. The ideal test involves controlled ablations, removing mechanism M and observing performance degradation; for modern neural networks with billions of parameters, identifying and cleanly ablating the “mechanism” responsible for a semantic function is non-trivial. Which parameters constitute the M that implements spatial reasoning? Attention heads? Specific layers? Furthermore, G2b’s requirement that M was acquired because it improved task success (not merely correlated with it) demands access to training dynamics and counterfactual training runs: computationally expensive and rarely documented.

For closed systems, we may be limited to observational proxies or behavioral tests providing only lower bounds on $\text{ACE}_E(M)$. G2b, then, is the most philosophically principled but practically challenging desideratum. We anticipate that many evaluations will satisfy only G2a (correlational faithfulness) without fully certifying G2b. Whether this suffices, whether correlation without etiology constitutes genuine grounding or merely reliable coincidence, remains a question we surface, but do not resolve here.

8.5 Indexing and context-sensitivity

Indexing by (k, t) provides a clean, analytical precision; it introduces decisions that affect reported parameters and inter-study comparability. The boundaries between meaning types are not always clean; many symbols simultaneously invoke multiple types. Consider “marriage”: it denotes a legal/social status, implies entailments about rights and obligations, and carries normative force about appropriate conduct. Evaluating *the* grounding of “marriage” requires specifying which t is under test, yet real communication blends these. Similarly, context specification remains underspecified: how finely should contexts be individuated? Should k track only task-domain (kitchen versus workshop), or also interlocutor identity, discourse history, and cultural background?

More fundamentally, our treatment of context-sensitivity by indexing parameters by k and allowing \mathcal{I}_k^t to vary captures some phenomena but undersells the context-dependence of natural language. Gradable adjectives (“tall”, “expensive”) shift continuously with comparison class in languages with positive degree, comparative degree, and superlative degree operate differently than languages with relative, for example; quantifier domain restriction tacitly limits “every student” to contextually relevant students; perspectival expressions and deixis (“local”, “nearby”, “here”, “there”) depend on speaker location and salience. Our framework can, in principle, handle these by k , but pushes complexity into context specification rather than addressing semantic flexibility intrinsically.

Vagueness sits uneasily with our metric-based approach, which assumes determinate $\mathcal{I}_k^t(\sigma)$ for each σ at each (k, t) . Supervaluationist or degree-theoretic treatments would require enriching meaning spaces with partial/fuzzy truth-values or probability distributions, which we do not develop here, and leave for future work. These considerations suggest the framework best applies to relatively stable, context-invariant meanings or carefully controlled (k, t) pairs. Whether this reflects a limitation of the framework or a genuine constraint on precise grounding evaluation is itself an open question.

8.6 Structural and temporal scope

Compositionality (G4) assumes typed grammatical constructors and a well-defined semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$. For systems without explicit grammatical structure (e.g., end-to-end neural networks), defining appropriate constructors becomes circular: we must impose structure to measure structure-preservation. More fundamentally, natural language exhibits both compositional and non-compositional constructions. Phrases like “kick the bucket” or “red herring” do not derive meanings from parts in the systematic way G4 demands. Our framework handles this via tolerance $\delta_{\text{comp}}^{k,t}$, treating perfect composition as a limiting behavior and idiomatic expressions as contributing to larger $\delta_{\text{comp}}^{k,t}$ values.

This treatment records deviation without explaining when non-compositionality is acceptable. Future work should distinguish *productive* composition (novel combinations interpreted systematically) from *lexicalized* idioms (stored whole units), and so model the gradient between them. Construction grammar and usage-based approaches suggest compositionality exists on a spectrum shaped by frequency and conventionalization [7, 26]. Our $\delta_{\text{comp}}^{k,t}$ collapses this spectrum into a single number; future refinements should track separate parameters for productivity on novel forms versus storage of conventional expressions.

Relatedly, our analysis shows that uniformly discrete systems (Corollary A.1) exhibit degenerate moduli: formally continuous; practically fragile. While this accurately characterizes brittleness under perturbations like typos (where single-character edits can flip meanings arbitrarily), it offers limited guidance for improvement.

Finally, our framework captures grounding *at a fixed point*, but does not address temporal dynamics. Language evolves through development (child acquisition, where meanings stabilize over years) and cultural change (semantic drift, where “awful” shifts from “awe-inspiring” to “terrible” over centuries, or “literally” acquires an intensifier use). Similarly, model grounding may shift: a model’s $\varepsilon_{\text{faith}}^{k,t}$ on medical terminology may decrease after domain adaptation. Static parameters cannot capture these trajectories, predict stability under continued learning, or distinguish robust grounding (maintained across distributional shifts) from brittle grounding (dependent on narrow training conditions). Extending the framework to track grounding profiles over time via time-indexed (k_τ, t) or by casting $\varepsilon_{\text{pres}}^{k,t}(\tau)$ as curves remains future work.

8.7 Application-specific adequacy

We argue throughout that partial grounding suffices for many applications, but provide no systematic theory for determining *which* parameters matter for *which* tasks. Heuristically: a dialogue system might tolerate higher $\delta_{\text{comp}}^{k,t}$ (accepting non-compositional phrases) but require low $\varepsilon_{\text{faith}}^{k,t}$ (accurate intent tracking); a robotic controller may exhibit the opposite pattern (tolerating intent ambiguity if spatial grounding is precise). The framework measures and reports, but does not prescribe task- or domain-specific requirements.

A complete engineering methodology in a logical next step from here, that would map task families to grounding profiles, specifying which desiderata are negotiable and which constitute hard constraints. This connects to ongoing work in AI safety and alignment: determining adequate grounding for high-stakes applications requires normative judgment about acceptable risk under semantic uncertainty. We view the framework as providing the measurement vocabulary; the adequacy thresholds remain domain-specific and, in many cases, contested.

9 Conclusion

We have recast the symbol grounding problem into measurement, and provisioned a language for discussion therein; \mathfrak{G} operationalizes grounding through four-plus-one desiderata, indexed by context (k) and meaning type (t), with explicit tolerances that enable assessment: preservation; faithfulness; robustness; compositionality; authenticity.

Model-theoretic semantics achieves compositionality (G4) but fails etiological warrant (G2b), aligned with Peregrin’s diagnosis: a theory of inference, not meaning. Pure symbolic systems satisfy structural requirements by stipulation, but lack the causality that ties symbols to their purported referents. Large language models are intermediate, achieving correlational faithfulness and local robustness through distributional learning while lacking the selection-for-success that characterizes genuine world-grounding. Natural language, shaped by evolutionary and developmental pressures across perceptual, inferential, and social dimensions, satisfies desiderata under authenticity.

Rather than asking whether LLMs “understand” or possess “real meaning,” we can access specific deficits: they achieve G2a (correlational fit) for linguistic tasks but fail G2b (etiological warrant) for embodied control; they exhibit local robustness (G3) under benign perturbations but lack global bounds under adversarial inputs. In this way, we advocate targeted improvements rather than wholesale dismissal or otherwise uncritical acceptance.

Implications follow: grounding is typed, and systems may be grounded in one mode while failing in others; the authenticity requirement (G0) blocks post-hoc stipulation, in which meanings must be earned through the agent’s causal mechanisms, in contrast to imposition by external analysts, which distinguishes genuine grounding from mere modeling; perfect grounding is neither necessary, nor, outside natural language, achieved: artificial systems can be useful with partial grounding, provided their limitations are explicit and their tolerances measured. We emphasize: instead of “Is this system grounded?”, a question that invites ideological posturing, we ask “What are its grounding parameters at (k, t) under threat model U ?”.

References

- [1] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.
- [2] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.
- [3] Trapit Bansal, Da-Cheng Juan, Sujith Ravi, and Andrew McCallum. A2n: Attending to neighbors for knowledge graph inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4387–4392, 2019.
- [4] Stéphanie Barbu, Aurélie Nardy, Jean-Pierre Chevrot, and Jacques Juhel. Language evaluation and use during early childhood: Adhesion to social norms or integration of environmental regularities? *Linguistics*, 51(2):381–411, 2013.
- [5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event, Canada, March 2021. Association for Computing Machinery.
- [6] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.
- [7] Hans C. Boas. Construction grammar in the twenty-first century. *English Language and Linguistics*, 11(3):569–585, 2007.
- [8] Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 4923–4931, 2021.
- [9] Robert Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1994.
- [10] Robert Brandom. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford University Press, 2010.
- [11] Rodney A. Brooks. Elephants don’t play chess. *Robotics and Autonomous Systems*, 6(1):3–15, 1990. Designing Autonomous Agents.
- [12] Tyler Burge. Individualism and the mental. *Midwest Studies in Philosophy*, 4(1):73–122, 1979.
- [13] Angelo Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, 7(2):139–151, 2010.
- [14] Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open*, 2:143–159, 2021.
- [15] Rémi Catellier, Samuel Vaiter, and Damien Garreau. On the robustness of text vectorizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3782–3814. PMLR, 2023.
- [16] Chin-Liang Chang and Richard Char-Tung Lee. *Symbolic logic and mechanical theorem proving*. Academic press, 2014.
- [17] Yunliang Chen, Yuqi Li, Xiaohui Huang, Yuewei Wang, Guishui Zhu, Geyong Min, and Jianxin Li. An explainable recommendation method for artificial intelligence of things based on reinforcement learning with knowledge graph inference. *IEEE Transactions on Consumer Electronics*, 2024.
- [18] Silvia Coradeschi, Amy Loutfi, and Britta Wrede. A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, 27:129–136, 2013.
- [19] Kaustubh Dhole, Ramraj Chandradevan, and Eugene Agichtein. Adversem: Adversarial robustness testing and training of llm-based groundedness evaluators via semantic structure manipulation. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 395–408, 2025.
- [20] Ezequiel A. Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. *Linguistic Bodies: The Continuity between Life and Language*. The MIT Press, 11 2018.
- [21] Mark Dingemanse and Nick J Enfield. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42, 2024.

- [22] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.
- [23] M. C. Escher. Drawing hands, 1948. Lithograph, 28.2 cm OE 33.2 cm.
- [24] James Fodor, Simon De Deyne, and Shinsuke Suzuki. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190, 03 2025.
- [25] James W Garson. *What logics mean: from proof theory to model-theoretic semantics*. Cambridge University Press, 2013.
- [26] A.E. Goldberg. *Constructions at Work: The Nature of Generalization in Language*. Oxford linguistics. Oxford University Press, 2006.
- [27] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.
- [28] Reto Gubelmann. Pragmatic norms are all you need – why the symbol grounding problem does not apply to LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [29] Jacqueline Harding. Operationalising representation in natural language processing. *British Journal for the Philosophy of Science*, 2023.
- [30] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.
- [31] Stevan Harnad. Language writ large: Llm, chatgpt, meaning and understanding. *Frontiers in Artificial Intelligence*, 7:1490698, 2024.
- [32] Christopher A Hennessy and Charles AE Goodhart. Goodhart’s law and machine learning: a structural perspective. *International Economic Review*, 64(3):1075–1086, 2023.
- [33] Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. Grounded language learning fast and slow. *CoRR*, abs/2009.01719, 2020.
- [34] Steve R Howell, Damian Jankowicz, and Suzanna Becker. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, 53(2):258–276, 2005.
- [35] Kristiina Jokinen. The need for grounding in llm-based dialogue systems. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge)@LREC-COLING-2024*, pages 45–52, 2024.
- [36] Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6523–6533, 2024.
- [37] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [38] Hiu Ting Lau and Arkaitz Zubiaga. Understanding the effects of human-written paraphrases in llm-generated text detection. *Natural Language Processing Journal*, 11:100151, 2025.
- [39] Jinu Lee, Qi Liu, Runzhi Ma, Vincent Han, Ziqi Wang, Heng Ji, and Julia Hockenmaier. Entailment-preserving first-order logic representations in natural language entailment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5729–5742, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [40] Hector J Levesque and Gerhard Lakemeyer. *The logic of knowledge bases*. Mit Press, 2001.
- [41] Stephen C Levinson. Yélf dnye and the theory of basic color terms. *Journal of linguistic anthropology*, 10(1):3–55, 2000.
- [42] M. Lewis and M. Mitchell. Evaluating the robustness of analogical reasoning in GPT models. *Transactions on Machine Learning Research*, 2025.

- [43] Yansheng Li, Song Ouyang, and Yongjun Zhang. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. *Knowledge-Based Systems*, 243:108469, 2022.
- [44] Yihong Liu, Raoyuan Zhao, Lena Altinger, Hinrich Schütze, and Michael A. Hedderich. Evaluating robustness of large language models against multilingual typographical errors, 2025.
- [45] David Manheim and Scott Garrabrant. Categorizing variants of goodhart’s law, 2019.
- [46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, arXiv:1301.3781, 2013. Published on 16 January 2013.
- [47] Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Representation Learning*, pages 94743–94765, 2025.
- [48] Dimitri Coelho Mollo and Raphaël Millièvre. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023.
- [49] Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*, 2021.
- [50] Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoenybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *arXiv preprint arXiv:2407.14679*, 2024.
- [51] Uzma Nawaz, Mufti Anees ur Rahaman, and Zubair Saeed. A review of neuro-symbolic ai integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, 26:200541, 2025.
- [52] Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041, 2023.
- [53] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [54] Jaroslav Peregrin. Language and its models: Is model theory a theory of semantics? *Nordic Journal of Philosophical Logic*, 2(1):1–23, 1997.
- [55] Francesco Periti and Stefano Montanelli. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–38, 2024.
- [56] Marla Perkins. Colors, sounds, rivers: Descriptive terms and informational priority in hobongan. *Journal of Borneo-Kalimantan*. Vol, 10:1, 2024.
- [57] Hilary Putnam. The meaning of ‘meaning’. *Minnesota Studies in the Philosophy of Science*, 7:131–193, 1975.
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
- [59] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(4):1–27, 2018.
- [60] Sara Rajaei and Mohammad Taher Pilehvar. How does fine-tuning affect the geometry of embedding space: A case study on isotropy, 2021.
- [61] Jacob Russin, Sam Whitman McGrath, Danielle J. Williams, and Lotem Elber-Dorozko. From frege to chatgpt: Compositionality in language, cognition, and deep neural networks, 2024.
- [62] Lenhart K. Schubert, Benjamin Van Durme, and Marzieh Bazrafshan. Entailment inference in a natural logic-like general reasoner. In *AAAI Fall Symposium: Commonsense Knowledge*, 2010.
- [63] John Searle. Minds, brains and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.
- [64] S Selva Birunda and R Kanniga Devi. A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pages 267–281, 2021.
- [65] N. E. Sharkey and S. A. Jackson. Three horns of the representational trilemma. In V. Honavar and L. Uhr, editors, *Symbol Processing and Connectionist Models for Artificial Intelligence and Cognitive Modeling: Steps towards Integration*, pages 155–189. Academic Press, Cambridge, MA, 1994.

- [66] Chen Shengyuan, Yunfeng Cai, Huang Fang, Xiao Huang, and Mingming Sun. Differentiable neuro-symbolic reasoning on large-scale knowledge graphs. *Advances in Neural Information Processing Systems*, 36:28139–28154, 2023.
- [67] Jihao Shi, Xiao Ding, Kai Xiong, Hengwei Zhao, Bing Qin, and Ting Liu. Natural logic at the core: Dynamic rewards for entailment tree generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17372–17382, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [68] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.
- [69] Vivian S Silva, André Freitas, and Siegfried Handschuh. Exploring knowledge graphs in an interpretable composite approach for text entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7023–7030, 2019.
- [70] Ayush Singh, Navpreet Singh, and Shubham Vatsal. Robustness of llms to perturbations in text. *arXiv preprint arXiv:2407.08989*, 2024.
- [71] Anders Søgaard. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33(1):33–54, 2023.
- [72] Luc Steels and P. Vogt. Grounding adaptive language games in robotic agents. In I. Harvey and P. Husbands, editors, *Proceedings of the 4th European Conference on Artificial Life*, pages 474–482, Cambridge, MA, 1997. The MIT Press. Language.
- [73] Anushka Swarup, Avanti Bhandarkar, Ronald Wilson, Tianyu Pan, and Damon Woodard. From syntax to semantics: Evaluating the impact of linguistic structures on llm-based information extraction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 36–48, 2025.
- [74] Pirkko T [otilde] ugu and Tiia Tulviste. References to social norms by preschool children and their linguistic expression. *European Journal of Developmental Psychology*, 7(2):249–264, 2010.
- [75] Mariarosaria Taddeo and Luciano Floridi. Solving the symbol grounding problem: a critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):419–445, 2005.
- [76] Praneeth Vadlapati. Investigating the impact of linguistic errors of prompts on llm accuracy. *ESP Journal of Engineering & Technology Advancements*, 3(2):144–147, 2023.
- [77] Francisco J. Varela, Eleanor Eleanor Rosch, and Evan Thompson. Symbols. In *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, 09 1991.
- [78] Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F. Chen. Resilience of large language models for noisy instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11939–11950. Association for Computational Linguistics, November 2024.
- [79] Yinwei Wei, Xiang Wang, Liqiang Nie, Shaoyu Li, Dingxian Wang, and Tat-Seng Chua. Causal inference for knowledge graph based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11153–11164, 2022.
- [80] Bodo Winter. Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, 36(10):960–967, 2014.
- [81] Bodo Winter and Morten H Christiansen. Robustness as a design feature of speech communication. In *The Evolution of Language*, pages 384–391. World Scientific, 2012.
- [82] Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability, 2024.
- [83] Da Yin, Tao Meng, and Kai-Wei Chang. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*, 2020.
- [84] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. Multimodal knowledge alignment with reinforcement learning, 2022.
- [85] Thomas Ede Zimmermann, C Maienborn, K von Heusinger, and P Portner. Model-theoretic semantics. *Semantics-Theories*, pages 181–231, 2019.

A Supplemental theorem material

Uniform continuity and a (global) modulus are equivalent formulations: a single monotone curve ω globally upper-bounds semantic change as a function of representation change. The tightest such curve is the minimal oscillation ω^* , defined purely by S and the two metrics.

Theorem A.1

[Modulus and uniform continuity] Fix (k, t) and let $S = \mathcal{A}_k^t \circ \Gamma : (\mathcal{R}, d_{\mathcal{R}}) \rightarrow (\mathcal{M}_k^t, d_{k,t})$. The following are equivalent:

1. S is uniformly continuous.
2. There exists a (global) modulus $\omega : [0, \infty) \rightarrow [0, \infty]$ that is monotone nondecreasing, satisfies $\omega(0) = 0$ and $\lim_{\varepsilon \downarrow 0} \omega(\varepsilon) = 0$, and

$$d_{k,t}(S(r), S(r')) \leq \omega(d_{\mathcal{R}}(r, r')) \quad \forall r, r' \in \mathcal{R}.$$

Moreover, the minimal oscillation function

$$\omega^*(\varepsilon) := \sup \{ d_{k,t}(S(r), S(r')) : d_{\mathcal{R}}(r, r') \leq \varepsilon \}$$

is such a modulus.

Proof: Let $\eta > 0$. By $\lim_{\varepsilon \downarrow 0} \omega(\varepsilon) = 0$, pick $\delta > 0$ with $\omega(\delta) < \eta$. If $d_{\mathcal{R}}(r, r') < \delta$, then by monotonicity of ω ,

$$d_{k,t}(S(r), S(r')) \leq \omega(d_{\mathcal{R}}(r, r')) \leq \omega(\delta) < \eta.$$

Thus S is uniformly continuous.

Now, define

$$\omega^*(\varepsilon) := \sup \{ d_{k,t}(S(r), S(r')) : d_{\mathcal{R}}(r, r') \leq \varepsilon \}.$$

For monotonicity, if $0 \leq \varepsilon_1 \leq \varepsilon_2$, then the set over which the supremum is taken for ε_1 is a subset of that for ε_2 , hence $\omega^*(\varepsilon_1) \leq \omega^*(\varepsilon_2)$.

Now for zero, the constraint $d_{\mathcal{R}}(r, r') \leq 0$ forces $d_{\mathcal{R}}(r, r') = 0$. In a (pseudo)metric, this implies $d_{k,t}(S(r), S(r')) = 0$; hence $\omega^*(0) = 0$.

For any r, r' , the pair belongs to the set with $\varepsilon = d_{\mathcal{R}}(r, r')$. Therefore

$$d_{k,t}(S(r), S(r')) \leq \omega^*(d_{\mathcal{R}}(r, r')).$$

Vanishing at zero follows. By uniform continuity, for any $\eta > 0$ there exists $\delta > 0$ such that $d_{\mathcal{R}}(r, r') < \delta$ implies that $d_{k,t}(S(r), S(r')) < \eta$. Then for every $\varepsilon \leq \delta$, each pair with $d_{\mathcal{R}}(r, r') \leq \varepsilon$ has image distance $< \eta$, so $\omega^*(\varepsilon) \leq \eta$. Hence $\lim_{\varepsilon \downarrow 0} \omega^*(\varepsilon) = 0$.

Thus ω^* is a modulus. □

Lemma A.1

[Minimality of ω^*] If ω is any modulus satisfying Theorem A.1, then for all $\varepsilon \geq 0$, $\omega^*(\varepsilon) \leq \omega(\varepsilon)$.

Proof: Fix ε and any r, r' with $d_{\mathcal{R}}(r, r') \leq \varepsilon$. By Theorem A.1 and monotonicity of ω ,

$$d_{k,t}(S(r), S(r')) \leq \omega(d_{\mathcal{R}}(r, r')) \leq \omega(\varepsilon).$$

Taking the supremum over all such pairs gives $\omega^*(\varepsilon) \leq \omega(\varepsilon)$. □

A uniformly discrete domain means that all distinct points in the representation space are separated by at least some fixed minimum distance; there are no pairs of representations that can get arbitrarily close to one another. Because of this gap, continuity becomes trivial: whenever two points are closer than δ_0 , they must, in fact, be the same point. As a result, any mapping $S = \mathcal{A}_k^t \circ \Gamma$ from such a domain to its meaning space is automatically uniformly continuous: small changes in representation cannot produce any change in meaning, because there simply are no smaller steps to take.

The modulus of continuity $\omega^*(\varepsilon)$ therefore looks like a step function: it is exactly zero for perturbations smaller than the minimal spacing δ_0 and bounded by the total range of possible semantic values for larger perturbations.

Corollary A.1

[Uniformly discrete domain] Assume $(\mathcal{R}, d_{\mathcal{R}})$ is uniformly discrete: there exists $\delta_0 > 0$ such that $d_{\mathcal{R}}(r, r') \geq \delta_0$ whenever $r \neq r'$. Then $S = \mathcal{A}_k^t \circ \Gamma$ is uniformly continuous. Moreover, the minimal oscillation function

$$\omega^*(\varepsilon) := \sup\{d_{k,t}(S(r), S(r')) : d_{\mathcal{R}}(r, r') \leq \varepsilon\}$$

satisfies $\omega^*(\varepsilon) = 0$ for every $\varepsilon < \delta_0$, and $\omega^*(\varepsilon) \leq \text{diam}(S(\mathcal{R}))$ for all $\varepsilon \geq \delta_0$. If the domain has finite diameter $D_{\mathcal{R}} := \sup_{r, r'} d_{\mathcal{R}}(r, r') < \infty$, then for every $\varepsilon \geq D_{\mathcal{R}}$ we have $\omega^*(\varepsilon) = \text{diam}(S(\mathcal{R}))$. In particular, discreteness alone yields a crude global Lipschitz bound whenever $\text{diam}(S(\mathcal{R})) < \infty$: for $r \neq r'$,

$$d_{k,t}(S(r), S(r')) \leq \text{diam}(S(\mathcal{R})) \leq \frac{\text{diam}(S(\mathcal{R}))}{\delta_0} d_{\mathcal{R}}(r, r'),$$

so S is L -Lipschitz with $L = \text{diam}(S(\mathcal{R}))/\delta_0$. Without further structure, no finite universal Lipschitz constant can be guaranteed when $\text{diam}(S(\mathcal{R})) = \infty$.

Proof: Uniform continuity is immediate: take $\delta = \delta_0$. If $d_{\mathcal{R}}(r, r') < \delta_0$ then $r = r'$ and hence $d_{k,t}(S(r), S(r')) = 0$; this meets the $\varepsilon - \delta$ condition. For the oscillation, when $\varepsilon < \delta_0$ the constraint $d_{\mathcal{R}}(r, r') \leq \varepsilon$ forces $r = r'$, so $\omega^*(\varepsilon) = 0$.

When $\varepsilon \geq \delta_0$, the supremum over admissible pairs can never exceed $\text{diam}(S(\mathcal{R}))$ by definition, which gives the upper bound for free. If, in addition, the domain diameter $D_{\mathcal{R}}$ is finite, then for any $\varepsilon \geq D_{\mathcal{R}}$ the constraint $d_{\mathcal{R}}(r, r') \leq \varepsilon$ becomes vacuous (every pair is allowed), so $\omega^*(\varepsilon)$ equals the diameter of the image.

The displayed Lipschitz bound follows by inserting the trivial inequality $d_{k,t}(S(r), S(r')) \leq \text{diam}(S(\mathcal{R}))$ and using $d_{\mathcal{R}}(r, r') \geq \delta_0$ for $r \neq r'$. \square

Corollary A.1 is why purely symbolic or discretized systems are formally smooth but are otherwise practically brittle. In a symbolic calculus with tokens or strings as representations, there is a hard jump between identity and complete difference, in which changing a single character or symbol can move the representation across the entire semantic space. There are no gradual transitions or intermediate cases, so the system satisfies continuity in a vacuous mathematical sense but fails to exhibit the graceful degradation we expect from robust grounding.

Remark. The step-function identity $\omega^*(\varepsilon) = \text{diam}(S(\mathcal{R}))$ for all $\varepsilon \geq \delta_0$ does not hold in general¹⁰: it requires that some pair (r, r') achieving (or approaching) the image diameter also satisfy $d_{\mathcal{R}}(r, r') \leq \delta_0$. Uniform discreteness alone does not ensure this.

Remark. If one assumes the discrete metric $d_{\mathcal{R}}(r, r') \in \{0, 1\}$ for $r \neq r'$, then indeed $\omega^*(\varepsilon) = 0$ for $\varepsilon < 1$ and $\omega^*(\varepsilon) = \text{diam}(S(\mathcal{R}))$ for $\varepsilon \geq 1$.

Remark. Counterexample (in the non-uniformly discrete domain) follows. Let $\mathcal{R} = \{0\} \cup \{1/n : n \in \mathbb{N}\} \subset \mathbb{R}$ with the Euclidean metric, and define $S(0) = 0$ while $S(1/n) = 1$ for all n . Then $r_n = 1/n \rightarrow 0$ but $d_{k,t}(S(r_n), S(0)) = 1$ for every n . Consequently, S fails to be uniformly continuous, and by Theorem A.1, no global modulus exists.

If atomic meanings are preserved exactly (no error on G1) and composition is exact (no error on G4), then the system's realized interpretation respects the grammar's constructors *on the nose*. Algebraically, $\mathcal{A}_k^t \Psi$ is a homomorphism from the term algebra to the semantic algebra, just as Frege and Montague intended.

Theorem A.2

[Exact homomorphism at $\varepsilon_{\text{pres}}^{k,t} = \delta_{\text{comp}}^{k,t} = 0$] Fix (k, t) . Let \mathcal{G} be a typed, free algebra of constructors (no equations beyond typing) over the atomic vocabulary $\Sigma_{\text{atom}} \subseteq \Sigma$, and let $\text{Terms}(\Sigma_{\text{atom}}, \mathcal{G})$ be the corresponding typed term algebra. Assume G1 holds with $\varepsilon_{\text{pres}}^{k,t} = 0$ on atoms, and typed G4 holds with $\delta_{\text{comp}}^{k,t} = 0$ on the closure of Σ_{atom} under \mathcal{G} . Writing $F := \mathcal{A}_k^t \Psi$, and identifying each well-typed term τ with its surface realization in Σ via \mathcal{G} (so $F(\tau)$ abbreviates $F(\text{surf}(\tau))$), the map

$$F : \langle \text{Terms}(\Sigma_{\text{atom}}, \mathcal{G}), \mathcal{G} \rangle \longrightarrow \langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$$

¹⁰Thank you to colleague, who wishes to remain anonymous, who articulated this point in workshop.

is a homomorphism: for every typed constructor f and well-typed tuple of terms (τ_1, \dots, τ_n) ,

$$F(f(\tau_1, \dots, \tau_n)) = f_k^{\mathcal{M}, t}(F(\tau_1), \dots, F(\tau_n)).$$

Proof: Assume G1 holds: with $\varepsilon_{\text{pres}}^{k, t} = 0$, $F(\sigma) = \mathcal{I}_k^t(\sigma)$ for all $\sigma \in \Sigma_{\text{atom}}$. We proceed by structural induction on $\tau \in \text{Terms}(\Sigma_{\text{atom}}, \mathcal{G})$. The base case is immediate.

For the induction step, suppose the claim holds for τ_1, \dots, τ_n and consider $\tau = f(\tau_1, \dots, \tau_n)$. By typed G4, with $\delta_{\text{comp}}^{k, t} = 0$,

$$F(f(\tau_1, \dots, \tau_n)) = f_k^{\mathcal{M}, t}(F(\tau_1), \dots, F(\tau_n)),$$

which is exactly the homomorphism property. \square

Homomorphisms out of a free term algebra are uniquely determined by their action on atoms. Exact G1 and G4, therefore, force $\mathcal{A}_k^t \Psi$ to coincide with the intended semantics on every well-typed expression.

Corollary A.2

[Coincidence with the intended semantics] For the typed grammar \mathcal{G} , let $\mathcal{I}_k^{t, \uparrow} : \text{Terms}(\Sigma_{\text{atom}}, \mathcal{G}) \rightarrow \mathcal{M}_k^t$ be the unique homomorphic extension of $\mathcal{I}_k^t|_{\Sigma_{\text{atom}}}$. Under Theorem A.2, $F = \mathcal{A}_k^t \Psi$ coincides with $\mathcal{I}_k^{t, \uparrow}$ on all terms:

$$\forall \tau \in \text{Terms}(\Sigma_{\text{atom}}, \mathcal{G}) : \mathcal{A}_k^t \Psi(\tau) = \mathcal{I}_k^{t, \uparrow}(\tau).$$

Proof:[Proof of Corollary A.2] By Theorem A.2, $F := \mathcal{A}_k^t \Psi$ is a homomorphism. By G1 with $\varepsilon_{\text{pres}}^{k, t} = 0$, F agrees with \mathcal{I}_k^t on atoms. By the universal property of the (typed) term algebra, there is a unique homomorphic extension $\mathcal{I}_k^{t, \uparrow}$ of \mathcal{I}_k^t , hence $F = \mathcal{I}_k^{t, \uparrow}$. \square

B Toy example

To demonstrate the application of the framework and clarify the independence of the desiderata, we detail here an extremely simplified toy example, a concrete evaluation of a toy agent in grid-world, of which we implemented and trained via policy gradient reinforcement learning. Code for this example may be found at the project repo.

B.1 Evaluation setup

We define the evaluation tuple $E = (k, t, U, P)$. The context k is a continuous 10×10 planar world; meaning type t is extensional, where meanings correspond to 2D coordinates; threat model U Gaussian noise injected into the representation space \mathcal{R} ; reference distribution P uniform over valid color-direction command pairs.

The agent’s vocabulary consists of atoms $\Sigma_{\text{atom}} = \{\text{RED}, \text{BLUE}, \text{NORTH}, \text{SOUTH}, \text{EAST}, \text{WEST}\}$. The intended interpretation \mathcal{I}_k^t maps colors to landmarks and directions to unit vectors:

$$\mathcal{I}_k^t(\text{RED}) = (8.0, 8.0), \quad \mathcal{I}_k^t(\text{NORTH}) = (0, 1).$$

The semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$ interprets the grammar’s composition constructor as vector addition:

$$\mathcal{I}_k^t(\text{RED NORTH}) = (8.0, 8.0) + (0, 1) = (8.0, 9.0).$$

B.2 Grounding architecture

The agent implements the grounding architecture \mathfrak{G} as follows:

- $\Phi : \Sigma^* \rightarrow \mathcal{R}$ embedding layer followed by a GRU, mapping token sequences to hidden states; the representation space is $\mathcal{R} = \mathbb{R}^{64}$ with Euclidean metric $d_{\mathcal{R}}$;
- $\Gamma : \mathcal{R} \rightarrow \mathcal{C}$ linear decoder mapping hidden states to predicted coordinates; concept space is $\mathcal{C} = \mathbb{R}^2$;
- $\mathcal{A}_k^t : \mathcal{C} \rightarrow \mathcal{M}_k^t$ the identity map; the agent navigates to its predicted coordinate, so $\mathcal{M}_k^t = \mathbb{R}^2$.

The end-to-end map is $\Psi = \Gamma \circ \Phi$, and the observable semantics is $S_{k,t} = \mathcal{A}_k^t \circ \Gamma$. Since \mathcal{A}_k^t is the identity, we write the agent’s output for input σ as $\mathcal{A}_k^t \Psi(\sigma)$.

The agent was trained using the classic REINFORCE algorithm over 3000 episodes, minimizing Euclidean distance to target coordinates. Training logs (an excerpt of which is shown in Table 6) confirm convergence from an initial loss of 5.23 to near-zero:

Table 6: Excerpt of training logs.

episode 0:	loss 5.234	(untrained)
episode 500:	loss 0.0002	
episode 2500:	loss 0.0033	(converged)

B.3 Grounding audit

Let us now walk through the grounding desiderata.

G0 (authenticity) The mapping mechanisms (Φ, Γ) are neural network weights acquired via reward-driven optimization; the system satisfies G0-strong: the semantic mappings are both implemented internally and acquired through a selection process (\mathcal{T} RL training) relevant to the task family.

G1 (preservation) We measure the deviation of atomic concepts from their targets. For the atom RED:

$$\mathcal{A}_k^t \Psi(\text{RED}) = (8.027, 7.987), \quad \mathcal{I}_k^t(\text{RED}) = (8.0, 8.0).$$

The preservation error is:

$$\varepsilon_{\text{pres}}^{k,t} = |(8.027, 7.987) - (8.0, 8.0)|_2 = 0.030.$$

On a 10×10 world, this represents drift of 0.3% of the total scale.

G2a (correlational faithfulness) For the composed expression RED NORTH:

$$\mathcal{A}_k^t \Psi(\text{RED NORTH}) = (8.041, 9.003), \quad \mathcal{I}_k^t(\text{RED NORTH}) = (8.0, 9.0).$$

The faithfulness error is:

$$\varepsilon_{\text{faith}}^{k,t} = |(8.041, 9.003) - (8.0, 9.0)|_2 = 0.041.$$

This is modestly larger than the atomic preservation error: accumulated drift across the two constituents.

G2b (etiological faithfulness) We identify the mechanism M responsible for processing directional modifiers (the GRU’s sequential integration of tokens beyond the first). We ablate M by restricting the encoder to process only the first token:

- M active: output $(8.041, 9.003)$, task success (distance < 0.5).
- M ablated: output $(8.027, 7.987)$, task failure.

The ablated output equals $\mathcal{A}_k^t \Psi(\text{RED})$ exactly: the agent navigates to the landmark, indeed, but ignores the directional modifier entirely. The average causal effect is $\text{ACE}_E(M) = 1.0$, the modifier-processing mechanism is strictly necessary for composed commands.

G3 (robustness) We inject Gaussian noise of magnitude $|\delta| = 0.5$ into the representation space \mathcal{R} (the 64-dimensional GRU hidden state) for the input RED:

- Unperturbed output: $(8.027, 7.987)$;
- Perturbed output: $(8.060, 7.905)$;
- Semantic drift: 0.088.

The local modulus satisfies $\omega_U^{k,t}(0.5) = 0.088$, yielding a ratio of approximately 0.18: less than one-fifth of the representational perturbation propagates to semantic output. This dampening follows from the dimensionality reduction through $\Gamma(\mathbb{R}^{64} \rightarrow \mathbb{R}^2)$ and the learned smoothness of the decoder. The behavior contrasts sharply with the degenerate step-function modulus of uniformly discrete symbolic systems (Corollary A.1), where any perturbation exceeding the minimal symbol spacing can induce arbitrary semantic change.

G4 (compositionality) We compare the agent’s composed output to the algebraic combination of its own atomic interpretations:

$$\begin{aligned}\mathcal{A}_k^t \Psi(\text{RED}) &= (8.027, 7.987), \\ \mathcal{A}_k^t \Psi(\text{NORTH}) &= (-0.009, 0.994).\end{aligned}$$

The agent’s learned representation of NORTH is nearly the true unit vector $(0, 1)$, with a small spurious negative x -component. If the agent were perfectly compositional:

$$f_k^{\mathcal{M},t}(\mathcal{A}_k^t \Psi(\text{RED}), \mathcal{A}_k^t \Psi(\text{NORTH})) = (8.027, 7.987) + (-0.009, 0.994) = (8.018, 8.981).$$

The actual output was $(8.041, 9.003)$. The compositionality deficit is:

$$\delta_{\text{comp}}^{k,t} = |(8.041, 9.003) - (8.018, 8.981)|_2 = 0.031.$$

B.4 Summary remarks

Taken together, we observe the following desiderata and values in Table 7.

Table 7: Grounding profile for gridworld agent.

Desideratum	Value	Interpretation
G0 (authenticity)	strong	learned via REINFORCE
G1 (preservation)	$\varepsilon_{\text{pres}}^{k,t} = 0.030$	atoms stable
G2a (faithfulness)	$\varepsilon_{\text{faith}}^{k,t} = 0.041$	composed outputs accurate
G2b (etiological)	$\text{ACE}_E(M) = 1.0$	strictly necessary mechanism
G3 (robustness)	$\omega_U^{k,t}(0.5) = 0.088$	strong dampening
G4 (compositionality)	$\delta_{\text{comp}}^{k,t} = 0.031$	highly systematic

First, $\varepsilon_{\text{faith}}^{k,t} > \varepsilon_{\text{pres}}^{k,t}$ (0.041 versus 0.030): the composed expression incurs modestly more error than the atoms alone; this is expected when composition accumulates independent atomic errors, though the increase here is slight.

Second, $\delta_{\text{comp}}^{k,t} \approx \varepsilon_{\text{pres}}^{k,t}$ (0.031 versus 0.030): the compositionality deficit is comparable to the atomic preservation error; the agent’s deviation from perfect composition is attributable primarily to imperfect atoms rather than to failures of the combinatory mechanism itself. The agent has learned to add correctly; it merely adds slightly wrong vectors.

Third, the robustness ratio $\omega_U^{k,t}(0.5)/0.5 \approx 0.18$ quantifies the degree to which the decoder Γ compresses high-dimensional representational noise into low-dimensional semantic output. This compression is a structural property of the architecture (projecting \mathbb{R}^{64} onto \mathbb{R}^2) combined with learned smoothness; a symbolic system with discrete \mathcal{R} would exhibit no such compression.

Finally, the ablation study (G2b) confirms that the measured success is not accidental: the sequential processing of modifiers is strictly necessary, and removing it reduces the agent to landmark-only navigation. The ACE of 1.0 establishes maximal etiological warrant for the modifier-integration mechanism.

Our grid-world shows that G2a and G4 are orthogonal, in that an agent could exhibit perfect compositionality ($\delta_{\text{comp}}^{k,t} = 0$) by consistently applying addition to incorrect atoms, even if both the parts and the whole were far from the ground truth; conversely, an agent could achieve perfect faithfulness on a phrase by memorizing the correct output without systematic combination of parts (high $\delta_{\text{comp}}^{k,t}$).

The agent demonstrates a different pattern: moderate accuracy with high systematicity. Its errors are primarily calibration errors in the atomic concepts. The grounding profile replaces the question “Does the agent understand?” with a quantified answer: it preserves atoms within 0.030 units, composes them with 0.031 units of systematic deviation, dampens representational noise by a factor of five, and its semantic mechanisms are causally necessary for task success. Whether this profile suffices depends on the application, of course: an assistant requiring 0.01-unit precision would reject this agent; an assistant tolerating 0.1-unit would accept it. The framework provides the measurements; adequacy thresholds remain domain-specific.

C Quick reference

An evaluation tuple $E = (k, t, U, P)$ fixes the context, meaning type, threat model, and reference distribution for grounding tests. The following subscripts are used throughout:

- Contexts k : k_{logic} (logical/analytic); k_{embodied} (embodied/task); k_{ling} (linguistic corpus), k_{vl} (multimodal).
- Meaning types t : ext (extensional/world-referential); inf (inferential/intensional/conceptual); soc (social-normative, reserved).
- Threat models U : U_{edit} (text edits); U_{nat} (naturalistic sensory noise); U_{ling} (textual/paraphrase noise); U_{vl} (image–text perturbations); U_{rel} (symbolic relation edits).
- Reference distributions P : P_{strings} (string distributions); $P_{\mathcal{R}}$ (representations, generic); $P_{\mathcal{R}}^{\text{ling}}$ (linguistic representations); $P_{\mathcal{R}}^{\text{vl}}$ (multimodal representations); $P_{\mathcal{R}}^{\text{rel}}$ (graphs/ontologies).

These subscripts specify the operative setting for each grounding mode: symbolic ($k_{\text{logic}}, \text{inf}, U_{\text{edit}}, P_{\text{strings}}$); referential ($k_{\text{embodied}}, \text{ext}, U_{\text{nat}}, P_{\mathcal{R}}$); vectorial-linguistic ($k_{\text{ling}}, \text{inf}, U_{\text{ling}}, P_{\mathcal{R}}^{\text{ling}}$); vectorial-multimodal ($k_{\text{vl}}, \text{ext}, U_{\text{vl}}, P_{\mathcal{R}}^{\text{vl}}$); relational ($k_{\text{logic}}, \text{inf}, U_{\text{rel}}, P_{\mathcal{R}}^{\text{rel}}$).

Table 8: Reference table of notation for grounding architecture and evaluation tuple.

Symbol	Meaning
\mathfrak{G}	grounding architecture with core structure of symbols, representations, concepts, meanings, mappings
Σ	alphabet of surface symbols (tokens, words, predicates, actions); $\Sigma_{\text{atom}} \subset \Sigma$ are atomic symbols
\mathcal{G}	grammar generating well-typed composites from Σ_{atom}
\mathcal{R}	representation space (latent or sensorimotor features); metric $d_{\mathcal{R}}$
\mathcal{C}	conceptual task-level substrate; metric $d_{\mathcal{C}}$
\mathcal{M}_k^t	meaning space for context k and meaning type t ; metric $d_{k,t}$; may carry semantic algebra $\langle \mathcal{M}_k^t, \mathcal{F}_k^t \rangle$
$\mathcal{I}_k^t : \Sigma \rightarrow \mathcal{M}_k^t$	intended interpretation; partial map extended homomorphically to composites as $\mathcal{I}_k^{t,\uparrow}$
$\Phi : \Sigma \rightarrow \mathcal{R}$	encoder from surface symbols to internal representations (tokenizer, perception, embedding)
$\Gamma : \mathcal{R} \rightarrow \mathcal{C}$	concept constructor, parser; builds structured conceptual states from representations (internal)
$\Psi = \Gamma \circ \Phi$	end-to-end symbol-to-concept map (context-agnostic)
$\mathcal{A}_k^t : \mathcal{C} \rightarrow \mathcal{M}_k^t$	alignment (semantics): projects concepts to the meaning space for the chosen (k, t)
$S_{k,t} = \mathcal{A}_k^t \circ \Gamma$	observable concept-to-meaning map used in evaluation
$E = (k, t, U, P)$	evaluation tuple: context, meaning type, threat model, reference distribution
k	evaluation context (e.g., $k_{\text{logic}}, k_{\text{embodied}}, k_{\text{ling}}, k_{\text{vl}}$).
t	meaning type (ext = extensional, inf = inferential/intensional, soc = social-normative)
U	threat model (family of perturbations); defines perturbation scale $\varepsilon(u)$ and robustness $\omega_U^{k,t}(\varepsilon)$
P	reference distribution over instances or representations used for expectation and robustness estimation
M	specific internal mechanism (e.g., module or parameter subset), causal contribution is tested in G2b
$\text{succ} : \mathcal{O} \rightarrow \{0, 1\}$	success predicate defining task achievement (used in $\text{ACE}_E(M)$)
$\text{ACE}_E(M)$	average causal effect of M under evaluation E ; check for etiological faithfulness G2b
$\varepsilon_{\text{pres}}^{k,t}$	preservation error on atomic meanings (G1)
$\varepsilon_{\text{faith}}^{k,t}$	correlational faithfulness error on composed meanings (G2a)
$\eta^{k,t}$	threshold for etiological faithfulness; minimum acceptable from $\text{ACE}_E(M)$
$\omega_U^{k,t}(\varepsilon)$	robustness modulus: upper bound on semantic deviation under perturbations of size ε
$\delta_{\text{comp}}^{k,t}$	compositional deviation: distance between composed meaning and meaning of parts (G4)
$\beta^{k,t}$	systematicity score on held-out (novel) compositional forms
$\text{GP}(\text{GA}; E)$	grounding profile at evaluation E : $[\varepsilon_{\text{pres}}^{k,t}, \varepsilon_{\text{faith}}^{k,t}, \text{ACE}_E(M), \omega_U^{k,t}(\cdot), \delta_{\text{comp}}^{k,t}, \beta^{k,t}]$
G0–G4	grounding desiderata: authenticity; preservation; faithfulness; robustness; compositionality
\mathcal{T}	training or evolutionary process under which mechanisms are acquired (for G0-strong and G2b)