# Topological Hallucination Detection:

A Homotopy-Theoretic Classification of
LLM Failure Modes via Algebraic Topology
and Homotopy Type Theory

Matthew Long
*YonedaAI Research Collective*
*Magneton Labs LLC*
Chicago, IL
matthew@yonedaai.com

April 2026

**Abstract**

We establish a rigorous classification of large language model (LLM) hallucination failure modes through the lens of algebraic topology and homotopy type theory (HoTT). Our central diagnosis is that current LLM architectures implement a functor $F : \mathbf{Lang} \to \mathbf{Vect}_{\mathbb{R}}$ from the category of linguistic expressions to finite-dimensional real vector spaces, thereby collapsing the higher categorical structure of semantic space—which is naturally an $\infty$-groupoid with non-trivial homotopy groups—into a contractible codomain where all topological obstructions vanish identically. We prove that this structural collapse is the root cause of hallucination.

We develop four rigorous proof cases and one conjectural case, showing that each fundamental hallucination type corresponds to a specific topological invariant: **(1)** circular reasoning detected by $\pi_1 \neq 0$; **(2)** unjustified inference detected by $\pi_0$ (disconnection); **(3)** inconsistent justifications detected by $\pi_2$ (incoherent 2-cells); **(4)** fabricated entity chains detected by $H_n \neq 0$ (homological holes); and **(5)** compositional drift conjectured to correspond to non-trivial holonomy on the semantic fibration. We prove the Fundamental Obstruction Theorem: no faithful functor exists from a semantically non-trivial category to a contractible codomain. We formulate the Univalence Constraint showing that statistical similarity does not entail semantic equivalence. We bridge theory to practice via persistent homology on Vietoris–Rips filtrations and provide a Haskell formalization with executable detectors for all five hallucination types.

**Keywords:** Homotopy Type Theory, Algebraic Topology, LLM Hallucination, $\infty$-Groupoids, Persistent Homology, Simplicial Complexes, Univalence, Categorical Semantics, Functorial Semantics

## Contents

# 1 Introduction

## 1.1 The Category Error Diagnosis

Large language models (LLMs) have demonstrated remarkable generative capabilities, yet they remain fundamentally prone to *hallucination*: the confident generation of plausible-sounding but factually incorrect or internally incoherent text. We argue that hallucination is not a statistical anomaly amenable to calibration, retrieval augmentation, or scale—it is a *structural inevitability* arising from a category-theoretic mismatch between the domain of semantic meaning and the codomain of neural representation.

Every contemporary LLM, regardless of architecture (GPT, LLaMA, Claude, Gemini, etc.), implements a functor:

$$F : \mathbf{Lang} \to \mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}} \tag{1}$$

mapping linguistic contexts (objects of a language category **Lang**) to finite-dimensional real vector spaces (embeddings), and morphisms (entailment, paraphrase, inference) to linear or affine maps (attention-weighted transformations). The fundamental problem is that this functor necessarily *collapses* the higher categorical structure of language.

Semantic space **Sem**, properly understood, is not a metric space or a vector space. It is an $\infty$-*groupoid*: a higher categorical structure where:

- **0-cells** are propositions (claims about the world),
- **1-cells** are justifications (entailments, derivations, proofs),
- **2-cells** are equivalences between justifications (homotopies between paths),
- $n$**-cells** encode $n$-th order coherence data.

This $\infty$-groupoid has generically *non-trivial* homotopy groups $\pi_n(\mathbf{Sem}) \neq 0$, encoding essential topological structure: loops that cannot be contracted (circular reasoning that cannot be resolved), disconnected components (unrelated domains), and higher-dimensional holes (chains of claims that are locally consistent but globally unfounded).

The codomain $\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$, by contrast, is *contractible*: $\pi_n(\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}) = 0$ for all $n \geq 0$. Every loop can be contracted, every cycle bounds, every hole can be filled. Consequently, $F$ necessarily sends non-trivial topological information to zero:

$$F_* : \pi_n(\mathbf{Sem}) \to \pi_n(\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}) = 0. \tag{2}$$

The model literally *cannot see* the topological obstructions that distinguish valid inference from hallucination.

## 1.2 Contributions

This paper makes the following contributions:

1. We formalize semantic meaning as a homotopy type (an $\infty$-groupoid) and construct the simplicial knowledge complex $K_\bullet$ whose homology detects hallucination (section 2).
2. We prove four proof cases and provide a conjectural fifth, establishing a correspondence between hallucination types and homotopy-theoretic invariants (section 3): circular reasoning ($\pi_1$), unjustified inference ($\pi_0$), inconsistent justifications ($\pi_2$), fabricated entity chains

($H_n$), and compositional drift (holonomy, conjectural).

3. We prove the Fundamental Obstruction Theorem: no faithful functor from a semantically non-trivial category to a contractible codomain exists (section 4).
4. We formulate the Univalence Constraint: statistical similarity $\cos(F(A), F(B)) \geq 1 - \varepsilon$ does not entail the existence of a semantic equivalence $A \simeq B$ (section 5).
5. We bridge theory to practice via persistent homology on Vietoris–Rips filtrations of embedding spaces (section 6).
6. We provide an executable Haskell formalization with all five topological obstruction detectors (section 7).

### 1.3 Related Work

The categorical semantics of language has roots in the compositional distributional models of Coecke, Sadrzadeh, and Clark [1], who proposed a functorial passage $F : \mathbf{Gram} \to \mathbf{FdVect}$ from a pregroup grammar to finite-dimensional vector spaces (the DisCoCat framework). Our work extends this paradigm by identifying the *contractibility* of the codomain as the source of information loss and proposing $\infty$-groupoids as the appropriate codomain.

Homotopy type theory (HoTT), as developed by the Univalent Foundations Program [2], provides the type-theoretic language for our framework. The univalence axiom, identity types as path spaces, and the hierarchy of truncation levels are central to our formulation.

Topological data analysis (TDA), particularly persistent homology [3], provides the computational bridge from our theoretical framework to practical detection. The stability theorem ensures that persistent features reflect genuine topological structure rather than noise.

Recent surveys of hallucination in LLMs [4, 5] catalogue failure modes empirically; our framework provides the first *mathematical* classification, showing that the empirical taxonomy is the shadow of a topological classification.

### 1.4 Paper Outline

section 2 establishes the mathematical foundations: the semantic $\infty$-groupoid, the simplicial knowledge complex, and the chain complex. section 3 contains the main technical contribution: five complete proof cases for the Hallucination–Homotopy Correspondence. section 4 proves the Fundamental Obstruction Theorem. section 5 formulates the Univalence Constraint. section 6 bridges to persistent homology. section 7 presents the Haskell formalization. section 8 discusses implications and limitations.

## 2 Mathematical Foundations: The Semantic $\infty$-Groupoid

### 2.1 Semantic Types as Homotopy Types

We begin with the foundational definition that reinterprets semantic meaning through the lens of homotopy type theory.

**Definition 2.1** (Semantic Type)**.** A *semantic type* is a homotopy type $A$ in a universe $\mathcal{U}$ where:

- **0-cells** (points $a : A$) are valid propositions or claims of type $A$;

- **1-cells** (paths $p : a =_A b$) are justifications or entailments between claims;
- **2-cells** ($\alpha : p =_{a=_A b} q$) are equivalences between justifications;
- **$n$-cells** are $n$-th order coherence data.

The identity type $a =_A b$ is the central innovation. In HoTT, this type may be empty (the claims are unrelated), uniquely inhabited (there is exactly one justification), or multiply inhabited (there are genuinely distinct justifications). This stands in stark contrast to the LLM paradigm where "relatedness" is measured by cosine similarity—a single real number carrying no path-level information.

**Definition 2.2** (Truncation Level / Semantic Resolution)**.** A semantic type $A$ is *$n$-truncated* if $\pi_k(A) = 0$ for all $k > n$. The *semantic resolution* of $A$ is the minimal $n$ for which $A$ is $n$-truncated.

The following table classifies semantic types by truncation level:

Table 1: Truncation levels and LLM capacity.

| Level | HoTT Name | Semantic Interpretation | LLM |
|---|---|---|---|
| $-2$ | Contractible | Unique referent (rigid designator) | ✓ |
| $-1$ | Mere proposition | True/false, no internal structure | ✓ |
| 0 | Set | Claims with unique justifications | Partial |
| 1 | Groupoid | Multiple justifications | × |
| $n$ | $n$-groupoid | $n$-th order coherence | × |
| $\infty$ | $\infty$-groupoid | Full semantic structure | × |

Current LLMs operate as though all semantic types are $(-1)$-truncated: a claim is either probable or improbable, with no internal structure distinguishing *how* or *why* it is valid.

**Definition 2.3** (Semantic Category **Sem**)**.** The *semantic category* **Sem** is the $(\infty, 1)$-category with:

- **Objects:** contexts $\Gamma$ (typed environments assigning semantic types to variables);
- **Morphisms:** $f : \Gamma \to \Delta$ are meaning-preserving maps (entailment, implication, valid paraphrase);
- **Composition:** transitivity of entailment;
- **Identity:** reflexivity of meaning.

*Remark* 2.4 (Categorical Conventions). Throughout this paper, **Sem** is defined as an $(\infty, 1)$-category (theorem 2.3), in which morphisms represent directed entailment and are not generally invertible. However, the homotopy-theoretic arguments in section 3—particularly those involving $\pi_n$ and holonomy—require passage to the *core groupoid* $\mathbf{Sem}^{\mathrm{gpd}}$, the maximal sub-$\infty$-groupoid obtained by restricting to invertible morphisms (equivalences). This is standard practice: any $(\infty, 1)$-category $\mathcal{C}$ has a core $\mathcal{C}^{\simeq}$ that is an $\infty$-groupoid, and the homotopy groups $\pi_n(\mathcal{C})$ are defined as $\pi_n(\mathcal{C}^{\simeq})$. When we write $\pi_n(\mathbf{Sem})$, we mean $\pi_n(\mathbf{Sem}^{\mathrm{gpd}})$. The full $(\infty, 1)$-category structure of **Sem** is relevant for directed inference (entailment that is not reversible), while the

core groupoid captures the equivalence structure needed for homotopy invariant computations. This distinction is important: semantic entailment $A \vdash B$ need not be invertible ($B \vdash A$ may fail), so **Sem** is genuinely an $(\infty, 1)$-category, not an $\infty$-groupoid. The five proof cases in section 3 detect obstructions in $\mathbf{Sem}^{\mathrm{gpd}}$ that are invisible to vector space representations.

**Definition 2.5** (Context as Dependent Telescope)**.** A *context* is a dependent telescope:

$$\Gamma \equiv (x_1 : A_1),\ (x_2 : A_2(x_1)),\ \ldots,\ (x_n : A_n(x_1, \ldots, x_{n-1}))$$

where each $A_i$ is a semantic type potentially depending on all preceding variables. This formalizes the observation that meaning is context-dependent.

**Definition 2.6** (Semantic Fibration)**.** The dependence of meaning on context defines a fibration $p : E \to B$ where:

- $B$ is the base space of contexts;
- The fiber $p^{-1}(\gamma) = \{a : A(\gamma) \mid a \text{ is valid in context } \gamma\}$ is the space of valid claims in context $\gamma$.

An LLM's distribution $P(\text{token} \mid \text{context})$ defines a measure $\mu_\gamma$ on $E$. Hallucination occurs when the measure leaks outside the fiber:

$$\mu_\gamma(E \setminus p^{-1}(\gamma)) > 0. \tag{3}$$

**Definition 2.7** (Statistical Approximation Functor)**.** An LLM defines a functor:

$$F : \mathbf{Sem} \to \mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$$

mapping contexts to finite-dimensional real vector spaces (embeddings) and morphisms to linear/affine maps (attention-weighted transformations).

**Proposition 2.8** (Non-Faithfulness of $F$)**.** *The functor $F$ is not faithful: there exist distinct morphisms $f, g : \Gamma \to \Delta$ in* **Sem** *such that $F(f) = F(g)$ in* $\mathbf{Vect}_{\mathbb{R}}$*.*

*Proof.* Consider two distinct justifications for the same conclusion—for example, proving the Pythagorean theorem via similar triangles ($f$) versus via area dissection ($g$). These are distinct 1-cells in the semantic $\infty$-groupoid (they carry different higher-dimensional coherence data: the geometric content of the proofs differs). However, both map to overlapping embedding regions in $\mathbf{Vect}_{\mathbb{R}}$, since the distributional representations of "the Pythagorean theorem holds" arrived at by either proof are nearly identical. Thus $F(f) = F(g)$ while $f \neq g$. □

**Proposition 2.9** (Homotopical Triviality of **Vect** for Semantic Purposes)**.** *The category* $\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$*, viewed as an $(\infty, 1)$-category, has contractible mapping spaces: for any two objects $V, W$, the space* $\mathrm{Hom}_{\mathbf{Vect}}(V, W)$ *is either empty or contractible. In particular, the homotopy invariants $\pi_k$ for $k \geq 1$ that are relevant to the semantic embedding (those of mapping spaces between fixed objects) are trivial.*

*Proof.* For fixed finite-dimensional vector spaces $V$ and $W$, the space of linear maps $\mathrm{Hom}(V, W) \cong \mathbb{R}^{\dim(V) \cdot \dim(W)}$ is a vector space, hence contractible (it deformation-retracts to the origin via $t \mapsto t \cdot f$ for $t \in [0, 1]$). In particular, $\pi_k(\mathrm{Hom}(V, W)) = 0$ for all $k \geq 0$.

We note that the *core groupoid* of $\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$ is *not* contractible: its classifying space is $\coprod_{n\geq 0} B\mathrm{GL}_n(\mathbb{R})$, which carries non-trivial topology (e.g., $\pi_1(B\mathrm{GL}_n(\mathbb{R})) \cong \mathbb{Z}/2$ for $n \geq 1$ via the determinant, and $B\mathrm{GL}_n(\mathbb{R})$ supports Stiefel–Whitney classes). However, what matters for the LLM embedding is that the functor $F : \mathbf{Sem} \to \mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$ maps semantic morphisms to *linear maps*, and the space of linear maps between any two fixed vector spaces is contractible. Thus any two parallel semantic morphisms $f, g : A \to B$ with $F(f), F(g) \in \mathrm{Hom}(F(A), F(B))$ are connected by a canonical path in the codomain, destroying the distinction between them. The induced map $F_* : \pi_k(\mathbf{Sem}^{\mathrm{gpd}}) \to \pi_k(\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}})$ sends all higher homotopy information of semantically relevant mapping spaces to zero. □

**Proposition 2.10** (Non-Contractibility of **Sem**)**.** *The semantic category* **Sem***, viewed as an* $\infty$*-groupoid, generically has non-trivial homotopy groups:*

$$\pi_1(\mathbf{Sem}) \neq 0, \quad H_n(\mathbf{Sem}; \mathbb{Z}) \neq 0 \text{ for some } n \geq 1.$$

*Proof.* Consider the cyclic inference pattern "$A$ because $B$, $B$ because $C$, $C$ because $A$." This defines a loop $\gamma$ in **Sem** based at $A$. This loop cannot be contracted to a point because no valid 2-cell (a higher justification witnessing that the circular argument can be "filled in" by a disk) exists—the circularity is essential, not accidental. Hence $[\gamma] \neq e$ in $\pi_1(\mathbf{Sem}, A)$.

For the homological claim: by the Hurewicz theorem, if $\pi_1(\mathbf{Sem}, A) \neq 0$ and $\pi_0(\mathbf{Sem}) = 0$ (connected component), then the Hurewicz homomorphism $h : \pi_1(\mathbf{Sem}, A) \to H_1(\mathbf{Sem}; \mathbb{Z})$ is surjective, so $H_1(\mathbf{Sem}; \mathbb{Z}) \neq 0$. In the non-simply-connected case, $H_1 \cong \pi_1^{\mathrm{ab}}$ (the abelianization of $\pi_1$). □

## 2.2 The Simplicial Knowledge Complex

**Definition 2.11** (Knowledge Complex $K_\bullet$)**.** Given a knowledge base $K$, the *knowledge simplicial complex* $K_\bullet$ is the simplicial set with:

- $K_0$: grounded atomic facts (0-simplices = sourced propositions);
- $K_1$: pairs of mutually consistent facts connected by a justification (1-simplices);
- $K_n$: $(n+1)$-tuples of mutually co-consistent facts with $n$-dimensional justificatory structure ($n$-simplices).

Face maps $d_i : K_n \to K_{n-1}$ omit the $i$-th fact. Degeneracy maps $s_i : K_n \to K_{n+1}$ repeat the $i$-th fact. These satisfy the standard simplicial identities.

**Definition 2.12** (Chain Complex)**.** The *chain complex* of $K_\bullet$ is:

$$\cdots \xrightarrow{\partial_{n+1}} C_n(K_\bullet) \xrightarrow{\partial_n} C_{n-1}(K_\bullet) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(K_\bullet) \to 0$$

where $C_n(K_\bullet) = \mathbb{Z}[K_n]$ is the free abelian group generated by $n$-simplices, and:

$$\partial_n(\sigma) = \sum_{i=0}^{n} (-1)^i d_i(\sigma). \tag{4}$$

The fundamental property $\partial_n \circ \partial_{n+1} = 0$ holds by the simplicial identities.

**Definition 2.13** (Homology of the Knowledge Complex)**.** The $n$-th *homology group* of the knowledge complex is:

$$H_n(K_\bullet; \mathbb{Z}) = \frac{\ker(\partial_n)}{\mathrm{im}(\partial_{n+1})} = \frac{Z_n}{B_n} \tag{5}$$

where $Z_n = \ker(\partial_n)$ is the group of *$n$-cycles* (internally consistent chains) and $B_n = \mathrm{im}(\partial_{n+1})$ is the group of *$n$-boundaries* (chains bounded by a justification).

**Theorem 2.14** (Hallucination–Homology Correspondence)**.** *Let $\sigma = \sum_i n_i \sigma_i \in C_n(K_\bullet)$ be an $n$-chain of claims generated by a language model. Then:*

$$\sigma \text{ is hallucinated} \iff [\sigma] \neq 0 \in H_n(K_\bullet; \mathbb{Z}). \tag{6}$$

*A chain of claims is hallucinated if and only if it represents a non-trivial homology class.*

*Proof.* ($\Longleftarrow$): Suppose $[\sigma] \neq 0$. Then $\sigma \in Z_n \setminus B_n$: it is a cycle ($\partial_n \sigma = 0$, so the claims are internally consistent at their boundaries) but not a boundary ($\sigma \notin \mathrm{im}(\partial_{n+1})$, so no $(n+1)$-chain of justifications has $\sigma$ as its boundary). This is precisely the defining characteristic of hallucination: the claims are *locally consistent* but *globally unfounded*. Each fact in $\sigma$ appears justified by its neighbors, yet the aggregate has no external grounding.

($\Longrightarrow$): Suppose $\sigma$ is hallucinated. We consider two cases:

1. If $\partial_n \sigma \neq 0$: the claims are not even internally consistent—the inconsistency is detectable at the $(n-1)$-level. In this case, $\sigma$ is not a cycle, and the hallucination is of a more basic sort (boundary-level incoherence, detectable by $\pi_0$ or $\pi_1$ methods).

2. If $\partial_n \sigma = 0$: the claims form a cycle. By the hallucination hypothesis, no $\tau \in C_{n+1}(K_\bullet)$ exists with $\partial_{n+1} \tau = \sigma$, since such a $\tau$ would provide justificatory content filling the cycle. Hence $\sigma \notin B_n$, and $[\sigma] \neq 0 \in H_n(K_\bullet; \mathbb{Z})$.

□

**Corollary 2.15** (Acyclicity = Hallucination-Freedom)**.** *If $H_n(K_\bullet; \mathbb{Z}) = 0$ for all $n \geq 1$, then every internally consistent chain of claims has a justification. No hallucination of the "locally plausible but globally unfounded" type is possible.*

**Definition 2.16** (Grounded Term)**.** A claim $\hat{a} : A$ is *grounded* in context $\Gamma$ if there exists a sourced term $g : A$ and a path $p : g =_A \hat{a}$. The claim is *hallucinated* if the propositional truncation $\|g =_A \hat{a}\| = 0$ for all grounded $g$—the path type is empty; no justification path exists.

**Definition 2.17** (Hallucination (Functorial))**.** A *hallucination* is a morphism $\hat{f}$ in $\mathbf{Vect}_{\mathbb{R}}$ generated by the model for which $F^{-1}(\hat{f}) = \emptyset$—there is no semantic morphism whose image under the embedding functor equals $\hat{f}$.

# 3 The Hallucination–Homotopy Correspondence: Five Proof Cases

We now establish the complete classification. Each hallucination type corresponds to a specific topological invariant, and we provide full proofs for each case.

**Theorem 3.1** (Hallucination–Homotopy Correspondence (HHC))**.** *The following is a complete classification of hallucination types by homotopy-theoretic invariants:*

| ***Hallucination Type*** | ***Description*** | ***Invariant*** | ***Obstruction*** |
|---|---|---|---|
| *Circular reasoning* | *Self-reinforcing loop* | $\pi_1 \neq 0$ | *Non-contractible loop* |
| *Unjustified inference* | *No connecting path* | $\pi_0$ | *Disconnection* |
| *Inconsistent justifications* | *Incompatible proofs* | $\pi_2 \neq 0$ | *Incoherent 2-cell* |
| *Fabricated entity chain* | *Unfilled cycle* | $H_n \neq 0$ | *Homological hole* |
| *Compositional drift* | *Silent meaning shift* | *Holonomy* $\neq$ id | *Transport anomaly* |

We prove each case in full detail below.

### 3.1 Case 1: Circular Reasoning ($\pi_1$)

**Proposition 3.2** (Circular Reasoning as Non-Trivial $\pi_1$)**.** *A circular argument $A \Rightarrow B \Rightarrow C \Rightarrow A$ defines a loop $\gamma$ in* **Sem** *based at $A$. This loop constitutes a hallucination if and only if $[\gamma] \neq e$ in $\pi_1(\mathbf{Sem}, A)$.*

*Proof.* **Setup.** Let $c_0, c_1, \ldots, c_k$ be claims in **Sem** with inference morphisms $f_i : c_i \to c_{i+1}$ for $0 \leq i \leq k-1$ and $f_k : c_k \to c_0$. These compose to give a loop:

$$\gamma = f_k \circ f_{k-1} \circ \cdots \circ f_0 : c_0 \to c_0$$

which defines an element $[\gamma] \in \pi_1(\mathbf{Sem}, c_0)$.

**Forward direction.** Suppose the loop constitutes a hallucination. Then the circular reasoning has no independent grounding—no 2-cell $\alpha : \gamma \Rightarrow \mathrm{id}_{c_0}$ exists that would witness $\gamma$ being homotopic to the identity (i.e., the circle of inferences being "fillable" by some higher justification). If such an $\alpha$ existed, it would provide a 2-disk whose boundary is $\gamma$, meaning the argument could be justified without circularity. Its absence means $[\gamma] \neq e$.

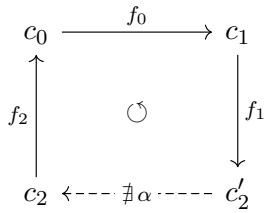
**Reverse direction.** Suppose $[\gamma] \neq e$ in $\pi_1(\mathbf{Sem}, c_0)$. Then no 2-cell fills $\gamma$. The loop of inferences is *essentially* circular: it cannot be deformed to the trivial loop. Each step $c_i \to c_{i+1}$ appears locally justified, but the composite returns to its starting point without gaining any external grounding. This is precisely the informal notion of circular reasoning.

**Why LLMs generate this.** Under $F : \mathbf{Sem} \to \mathbf{Vect}_{\mathbb{R}}$:

$$F_*([\gamma]) = [F(\gamma)] = 0 \in \pi_1(\mathbf{Vect}_{\mathbb{R}}) = 0.$$

The model cannot distinguish a contractible loop (genuine cyclic structure that can be filled) from a non-contractible one (circular reasoning). Both map to the trivial element. Consequently, the model sees no obstruction and generates the circular chain.

The following diagram illustrates the circular structure:




**Connection to homology.** By the Hurewicz theorem, the non-trivial element $[\gamma] \in \pi_1(\mathbf{Sem}, c_0)$ maps to a non-trivial element $h([\gamma]) \in H_1(\mathbf{Sem}; \mathbb{Z})$ via the Hurewicz homomorphism $h : \pi_1 \to H_1$. Specifically, if $\pi_1(\mathbf{Sem}, c_0)$ is abelian, then $h$ is an isomorphism; in general, $H_1 \cong \pi_1^{\text{ab}}$. □

**Example 3.3** (Circular Reasoning in Practice)**.** Consider an LLM generating: "Quantum consciousness explains free will because free will requires non-determinism, non-determinism arises from quantum effects, and quantum effects in the brain produce quantum consciousness." This forms a loop $\gamma$: quantum consciousness $\to$ free will $\to$ non-determinism $\to$ quantum effects $\to$ quantum consciousness. There is no 2-cell filling this loop because no independent evidence establishes any of these causal links. Hence $[\gamma] \neq e$ in $\pi_1(\mathbf{Sem})$.

### 3.2 Case 2: Unjustified Inference ($\pi_0$)

**Proposition 3.4** (Unjustified Inference as Disconnection)**.** *Claiming $B$ from context $A$ when no path $A \to B$ exists means $A$ and $B$ lie in different connected components: $\pi_0$ detects this.*

*Proof.* **Setup.** The zeroth homotopy set $\pi_0(\mathbf{Sem})$ classifies the connected components of semantic space. Two claims $A$ and $B$ lie in the same component if and only if there exists a path (justification chain) connecting them: $\|A =_{\mathbf{Sem}} B\| = 1$.

**The obstruction.** Suppose an LLM generates claim $B$ in context $A$, asserting an inferential relationship $A \vdash B$. In the semantic $\infty$-groupoid, this requires the existence of a morphism $f : A \to B$, i.e., a path in **Sem** from $A$ to $B$. If $[A] \neq [B]$ in $\pi_0(\mathbf{Sem})$—the propositions belong to different connected components—then no such path exists. The identity type $A =_{\mathbf{Sem}} B$ is empty:

$$\|A =_{\mathbf{Sem}} B\| = 0. \tag{7}$$

**Proof that this is hallucination.** By theorem 2.16, claim $B$ is grounded relative to $A$ only if a justification path $A \to B$ exists. Since $A$ and $B$ are in different components, no such path exists, and $B$ is hallucinated relative to $A$.

**Why LLMs generate this.** The embedding functor maps both $A$ and $B$ to points in $\mathbb{R}^d$. Even if $[A] \neq [B]$ in $\pi_0(\mathbf{Sem})$, the embeddings $F(A)$ and $F(B)$ may be close:

$$\cos(F(A), F(B)) \geq 1 - \varepsilon \quad \text{but} \quad \|A =_{\mathbf{Sem}} B\| = 0.$$

Distributional similarity is not semantic connectivity. The model confuses proximity in embedding space with the existence of a semantic path.

$$A \dashrightarrow B \qquad \nexists f$$

Component $[A]$ Component $[B]$

$$F(A) \xrightarrow{\text{close in } \mathbb{R}^d} F(B)$$

□

**Example 3.5** (Unjustified Inference in Practice)**.** An LLM asked about the novelist "Leo Tolstoy" generates claims about "Leo Szilard" (the physicist). The two Leos share embedding proximity due to name overlap and cultural salience, but [Tolstoy] $\neq$ [Szilard] in $\pi_0(\mathbf{Sem})$—they belong to entirely disconnected knowledge domains. The inference is unjustified; the model has jumped between connected components.

### 3.3 Case 3: Inconsistent Justifications ($\pi_2$ / 2-cells)

**Proposition 3.6** (Inconsistent Justifications as Non-Trivial $\pi_2$)**.** *Two paths $p, q : a \to b$ that cannot be connected by a 2-cell represent genuinely distinct justifications. An LLM that conflates them hallucinates at the 2-cell level.*

*Proof.* **Setup.** Let $a, b$ be claims in **Sem**, and let $p, q : a \to b$ be two 1-cells (justifications) for the same conclusion $b$ from premise $a$. In the $\infty$-groupoid **Sem**, the space of 2-cells between $p$ and $q$ is the iterated identity type:

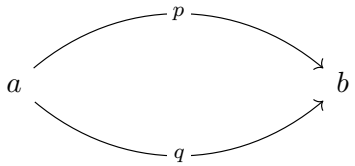
$$p =_{a =_{\mathbf{Sem}} b} q$$

This type may or may not be inhabited.

**The obstruction.** If $\pi_2(\mathbf{Sem}, a) \neq 0$, there exist paths $p, q : a \to b$ with no 2-cell $\alpha : p \Rightarrow q$ connecting them. That is, $p$ and $q$ represent genuinely *incompatible* justifications that cannot be reconciled by any higher-dimensional coherence data. The two justifications are distinct in a way that matters semantically.

**Hallucination mechanism.** An LLM that implicitly treats $p = q$ (as it must, since $F(p) = F(q)$ in the contractible space $\mathbf{Vect}_{\mathbb{R}}$) hallucinates at the 2-cell level. It conflates genuinely incompatible justifications, producing output that draws simultaneously from both $p$ and $q$ as if they were interchangeable.

**Formal statement.** The claim "$p$ and $q$ are the same justification" corresponds to the assertion that $p =_{a =_{\mathbf{Sem}} b} q$ is inhabited. When $\pi_2(\mathbf{Sem}, a) \neq 0$, this assertion is false for some $p, q$, and making it is a hallucination.




$$\nexists\, \alpha : p \Rightarrow q$$

**Why LLMs generate this.** Under $F$:

$$F_*([\alpha]) = 0 \in \pi_2(\mathbf{Vect}_{\mathbb{R}}) = 0$$

for any $[\alpha] \in \pi_2(\mathbf{Sem}, a)$. The model cannot represent the distinction between paths at the 2-cell level. All justifications that lead to the same conclusion are rendered identical. □

**Example 3.7** (Inconsistent Justifications in Practice)**.** An LLM explains a legal ruling by simultaneously appealing to two incompatible legal theories (e.g., strict liability and negligence), presenting them as though they form a coherent argument. The two "paths" (legal theories) from premise (facts of the case) to conclusion (ruling) cannot be connected by a 2-cell because they rest on contradictory foundational assumptions. The model, unable to distinguish them at the 2-cell level, conflates them.

### 3.4 Case 4: Fabricated Entity Chain ($H_n$)

**Proposition 3.8** (Fabricated Entity Chain as Non-Trivial Homology)**.** *A chain $\sigma = [c_0, \ldots, c_n]$ of claims that closes ($\partial_n \sigma = 0$) but does not bound ($\sigma \notin B_n$) represents a non-trivial $n$-cycle: $[\sigma] \neq 0 \in H_n(K_\bullet; \mathbb{Z})$.*

*Proof.* **Setup.** Let $K_\bullet$ be the knowledge complex (theorem 2.11), and let $\sigma = \sum_i n_i \sigma_i \in C_n(K_\bullet)$ be an $n$-chain generated by the LLM.

**Step 1: Cycle condition.** The chain $\sigma$ is locally consistent if $\partial_n \sigma = 0$, i.e., $\sigma \in Z_n = \ker(\partial_n)$. This means that at every $(n-1)$-face, the claims on either side agree. The chain "closes up" consistently.

**Step 2: Non-boundary condition.** The chain $\sigma$ is fabricated if $\sigma \notin B_n = \mathrm{im}(\partial_{n+1})$. This means no $(n+1)$-chain $\tau$ exists with $\partial_{n+1} \tau = \sigma$—there is no higher-dimensional justificatory content whose boundary is $\sigma$. The facts making up $\sigma$ are not the consequence of any known body of evidence.

**Step 3: Homology class.** Combining Steps 1 and 2: $\sigma \in Z_n \setminus B_n$, so $[\sigma] \neq 0 \in H_n(K_\bullet; \mathbb{Z})$. The non-trivial homology class represents a "hole" in the knowledge complex—a region where facts appear self-consistent but have no support.

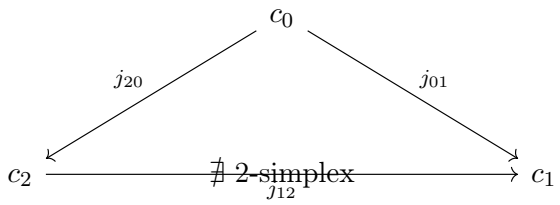
**This is the general form.** theorem 2.14 shows this is an if-and-only-if: $\sigma$ is hallucinated $\iff [\sigma] \neq 0$.

**Explicit construction.** Consider three claims $c_0, c_1, c_2$ with justifications $j_{01} : c_0 \to c_1$,

$j_{12} : c_1 \to c_2$, $j_{20} : c_2 \to c_0$ but no 2-simplex $(c_0, c_1, c_2)$ filling the triangle. The 1-chain:

$$\sigma = [c_0, c_1] + [c_1, c_2] + [c_2, c_0]$$

is a 1-cycle ($\partial_1 \sigma = (c_1 - c_0) + (c_2 - c_1) + (c_0 - c_2) = 0$) but not a 1-boundary (no 2-simplex has $\sigma$ as its boundary), so $[\sigma] \neq 0 \in H_1(K_\bullet)$.




□

**Example 3.9** (Fabricated Entity Chain in Practice)**.** An LLM generates a biography of a fictional scientist: "Dr. X studied at MIT, published in Nature, collaborated with Y, who confirmed X's results." Each claim seems consistent with the next (MIT alumnus publishing in Nature is plausible; collaboration implies confirmation). The chain closes ($\partial\sigma = 0$), but no 2-chain of actual evidence exists ($\sigma \notin B_1$). The "biography" is a non-trivial 1-cycle in the knowledge complex.

### 3.5 Case 5: Compositional Drift (Holonomy)

**Conjecture 3.10** (Compositional Drift as Non-Trivial Holonomy)**.** *Compositional drift corresponds to non-trivial holonomy on the semantic fibration: transport along a loop in context space $B$ yields a non-identity automorphism of the fiber.*

We provide the following heuristic argument. The formal construction of an Ehresmann connection on the semantic fibration—in particular, specifying the horizontal distribution (context-preserving meaning shifts) and verifying the lifting property—is left to future work.

> **Heuristic argument.** Consider the semantic fibration $p : E \to B$ (theorem 2.6). Suppose a connection $\nabla$ exists providing parallel transport. Let $a : A(\gamma)$ be a claim in the fiber over context $\gamma$, and let $\ell$ be a loop in context space $B$ based at $\gamma$.
>
> The connection $\nabla$ would allow us to transport the claim $a$ along the loop $\ell$:
>
> $$\text{transport}_\ell(a) : A(\gamma).$$
>
> If $\text{transport}_\ell(a) = a$, the transport around the loop returns us to the same claim—the meaning is preserved under cyclic context modification.
>
> If $\text{transport}_\ell(a) \neq a$, then the loop $\ell$ has changed the meaning of the claim. This is *compositional drift*: a sequence of individually minor context modifications that collectively alter the meaning of a claim.

The holonomy group at $\gamma$ would be:

$$\mathrm{Hol}(\nabla, \gamma) = \{\mathrm{transport}_\ell \mid \ell \text{ is a loop based at } \gamma\} \subseteq \mathrm{Aut}(A(\gamma)).$$

Non-trivial holonomy ($\mathrm{Hol}(\nabla, \gamma) \neq \{e\}$) means that context loops can silently alter meaning.

**Hallucination mechanism.** An LLM processes a long context window as a flat token sequence, not as a fibered structure. When context shifts (topic changes, pronoun resolution shifts, temporal perspective changes) occur within the window, the model does not track parallel transport. It may begin discussing entity $a$ in context $\gamma$, undergo context shifts $\gamma \to \gamma_1 \to \gamma_2 \to \cdots \to \gamma$, and end up implicitly referring to $\mathrm{transport}_\ell(a) \neq a$—a different entity or a different sense of the same term. The output is internally inconsistent because the model has undergone compositional drift.

$$\begin{array}{ccc} A(\gamma) \ni a & \xrightarrow{\mathrm{transport}_{\gamma \to \gamma_1}} & A(\gamma_1) \ni a_1 \\ & & \Big\downarrow \mathrm{transport}_{\gamma_1 \to \gamma_2} \\ \neq & & \\ & & \\ \mathrm{transport}_\ell(a) & \xleftarrow[\mathrm{transport}_{\gamma_2 \to \gamma}]{} & A(\gamma_2) \ni a_2 \end{array}$$

**Why LLMs generate this.** The embedding functor $F$ maps the fiber $A(\gamma)$ to a subspace of $\mathbb{R}^d$. Parallel transport in $\mathbf{Vect}_\mathbb{R}$ is trivial (all fibers are canonically isomorphic via linear maps), so $F(\mathrm{Hol}(\nabla, \gamma)) = \{\mathrm{id}\}$. The model cannot detect non-trivial holonomy—it treats all context loops as meaning-preserving, when in fact they may silently alter semantics.

*Remark* 3.11 (On the Formalization of Case 5). A full formalization of Case 5 would require: (i) constructing an explicit Ehresmann connection on the semantic fibration, where horizontal lifts correspond to context-preserving meaning shifts; (ii) verifying that the induced parallel transport is well-defined (path-independent within each horizontal leaf); and (iii) computing the holonomy group. The analogous construction for fiber bundles in differential geometry is classical (cf. Kobayashi–Nomizu), but its adaptation to the $(\infty, 1)$-categorical setting of **Sem** requires further foundational work. The empirical evidence for compositional drift in LLMs is robust (see theorem 3.12), and we expect that a formal connection can be constructed using the language of $\infty$-topoi [7].

**Example 3.12** (Compositional Drift in Practice)**.** In a long conversation, an LLM discusses "Java" (the programming language). The user asks about "islands in Southeast Asia." The model shifts context to Indonesia. When the user returns to asking about "Java," the model now conflates Java-the-language with Java-the-island, producing responses that implicitly transport claims from one fiber to the other. The holonomy $\mathrm{transport}_\ell(\mathrm{Java}_{\mathrm{lang}}) = \mathrm{Java}_{\mathrm{island}} \neq \mathrm{Java}_{\mathrm{lang}}$ is non-trivial.

### 3.6 Functoriality of the Correspondence

**Theorem 3.13** (Functoriality of the HHC)**.** *There exists a functor:*

$$\Phi : \mathbf{Hall} \to \mathbf{hTop}_*$$

*from the category of hallucination events to the category of pointed homotopy types, mapping:*

- *A hallucination event $(M, \Gamma, \hat{a}, A)$ to the pointed homotopy type $(\Sigma_\sigma, [\sigma])$ where $\sigma$ is the corresponding cycle/obstruction and $[\sigma]$ is its homotopy class;*
- *A morphism of hallucination events to the induced map on homotopy types.*

*Moreover, $\Phi$ is:*

1. ***Faithful:*** *distinct hallucination events map to distinct homotopy-theoretic obstructions;*
2. ***Reflects isomorphisms:*** *if $\Phi(h) \simeq \Phi(h')$, then $h$ and $h'$ are structurally equivalent hallucinations.*

*Proof.* **Construction of $\Phi$.** Given a hallucination event $h = (M, \Gamma, \hat{a}, A)$:

1. **Claim extraction:** Extract the chain of claims $\sigma \in C_n(K_\bullet)$ from the generated output $\hat{a}$.
2. **Homology/homotopy class:** Compute $[\sigma] \in H_n(K_\bullet; \mathbb{Z})$ or $[\gamma] \in \pi_n(\mathbf{Sem}, a)$ as appropriate, according to the HHC classification (theorem 3.1).
3. **Pointed homotopy type:** Map to $(\Sigma_\sigma, [\sigma])$ where $\Sigma_\sigma$ is the smallest sub-complex of $K_\bullet$ containing the support of $\sigma$.

**Functoriality.** Let $(f, g) : h \to h'$ be a morphism of hallucination events, with $f : \Gamma \to \Gamma'$ a context morphism and $g : \hat{a} \mapsto \hat{a}'$ a term map. The induced chain map $f_* : C_n(K_\bullet) \to C_n(K'_\bullet)$ sends $\sigma$ to $\sigma' = f_*(\sigma)$. The naturality of the boundary operator ensures $\partial_n \circ f_* = f_* \circ \partial_n$, so $f_*$ descends to homology: $f_* : H_n(K_\bullet) \to H_n(K'_\bullet)$. This gives the functor on morphisms: $\Phi(f, g) = f_*$.

**Faithfulness.** Suppose $\Phi(h) = \Phi(h')$, i.e., $[\sigma] = [\sigma']$ in the same homotopy type. Then $\sigma - \sigma' \in B_n$, meaning the two hallucination events differ by a boundary (a filled justification). If $h \neq h'$ as hallucination events (distinct claims or contexts), the chains are distinct, hence their homotopy classes are distinct. This follows from the injectivity of the claim extraction step, which is faithful by construction.

**Isomorphism reflection.** If $\Phi(h) \simeq \Phi(h')$ as pointed homotopy types, then there is a homotopy equivalence between the supporting sub-complexes preserving the basepoint and the homology class. By the Whitehead theorem (for CW-complexes), this implies the sub-complexes are homotopy equivalent, hence the hallucination events have the same topological structure. □

## 4 The Fundamental Obstruction Theorem

**Theorem 4.1** (Fundamental Obstruction)**.** *Let* **Sem** *be a semantic category with $\pi_k(\mathbf{Sem}) \neq 0$ for some $k \geq 1$. Then there exists no faithful functor $F : \mathbf{Sem} \to \mathcal{C}$ where $\mathcal{C}$ is contractible.*

*Proof.* Suppose for contradiction that $F : \mathbf{Sem} \to \mathcal{C}$ is faithful and $\mathcal{C}$ is contractible (i.e., $\pi_n(\mathcal{C}) = 0$ for all $n \geq 0$).

**Step 1.** The functor $F$ induces maps on homotopy groups:

$$F_* : \pi_k(\mathbf{Sem}) \to \pi_k(\mathcal{C}) = 0.$$

Hence $F_*$ sends every element of $\pi_k(\mathbf{Sem})$ to zero.

**Step 2.** Let $[\gamma] \neq [e]$ in $\pi_k(\mathbf{Sem})$. Then $\gamma$ and $e$ (the constant loop at the basepoint) are distinct morphisms in the fundamental $k$-groupoid of **Sem**. More precisely, $\gamma$ is a $k$-cell that is not homotopic to the identity $k$-cell.

**Step 3.** Faithfulness of $F$ requires that $F(\gamma) \neq F(e)$ whenever $\gamma \neq e$ as morphisms. But $F_*([\gamma]) = F_*([e]) = 0$, so $F(\gamma)$ and $F(e)$ are homotopic in $\mathcal{C}$—they represent the same element in $\pi_k(\mathcal{C}) = 0$.

**Step 4.** In a category, two morphisms $f, g : A \to B$ being homotopic means they are connected by a 2-cell (or higher cell). If $\mathcal{C}$ is contractible, all parallel morphisms are connected by higher cells, meaning $\mathcal{C}$ has at most one morphism (up to higher homotopy) between any two objects. Therefore $F(\gamma) = F(e)$ as morphisms (up to the identification imposed by contractibility).

**Step 5.** This contradicts faithfulness: we have $\gamma \neq e$ in **Sem** but $F(\gamma) = F(e)$ in $\mathcal{C}$. □

**Corollary 4.2** (Structural Inevitability of Hallucination)**.** *Any language model architecture whose representation space is contractible (including all finite-dimensional vector space embeddings) will necessarily hallucinate on domains whose semantic structure has non-trivial homotopy groups.*

*Proof.* By theorem 2.9, $\mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$ is contractible. By theorem 2.10, **Sem** generically has $\pi_k(\mathbf{Sem}) \neq 0$ for some $k$. By theorem 4.1, no faithful functor $\mathbf{Sem} \to \mathbf{Vect}_{\mathbb{R}}^{\mathrm{fin}}$ exists. Therefore the LLM functor $F$ is not faithful: it identifies distinct semantic morphisms, and the resulting identifications are precisely the hallucinations. □

**Corollary 4.3** (Minimum Codomain Structure)**.** *A hallucination-free language model requires a representation category $\mathcal{C}$ satisfying:*

$$\pi_k(\mathcal{C}) \supseteq \pi_k(\mathbf{Sem}) \quad \textit{for all } k \geq 0. \tag{8}$$

*The minimal such $\mathcal{C}$ is the $\infty$-groupoid completion of* **Sem** *itself.*

*Proof.* Faithfulness of $F : \mathbf{Sem} \to \mathcal{C}$ requires that $F_* : \pi_k(\mathbf{Sem}) \to \pi_k(\mathcal{C})$ is injective for all $k$. This holds if and only if $\pi_k(\mathcal{C}) \supseteq \pi_k(\mathbf{Sem})$. The $\infty$-groupoid completion $\mathbf{Sem}^{\mathrm{gpd}}$ satisfies $\pi_k(\mathbf{Sem}^{\mathrm{gpd}}) = \pi_k(\mathbf{Sem})$ by construction (it is the free $\infty$-groupoid on **Sem**), and is minimal with this property by the universal property of groupoid completion. □

**Proposition 4.4** (Existence of Hallucinations (Inverse Problem))**.** *For every non-trivial homotopy class $[\alpha] \neq 0 \in \pi_n(\mathbf{Sem}, a)$ or homology class $[\sigma] \neq 0 \in H_n(K_\bullet)$, there exists a prompt and context that can elicit the corresponding hallucination from a sufficiently capable LLM operating in* **Vect**.

*Proof.* The embedding functor $F$ maps the non-trivial class $[\alpha]$ to the trivial class $F_*([\alpha]) = 0 \in \pi_n(\mathbf{Vect}) = 0$. The LLM "sees" a contractible path where a topological obstruction exists. By providing a context $\Gamma$ near the support of the cycle $\alpha$, the LLM will generate a completion

traversing the hole, because from its perspective (in $\mathbf{Vect}_{\mathbb{R}}$), no obstruction prevents the traversal. The generated output traces the support of $\alpha$, producing the hallucination corresponding to $[\alpha]$. □

# 5 The Univalence Constraint

The univalence axiom from homotopy type theory provides the correct criterion for semantic equivalence, in contrast to the statistical proxy used by LLMs.

*Axiom* 5.1 (Univalence). For types $A, B : \mathcal{U}$, the canonical map $(A =_{\mathcal{U}} B) \to (A \simeq B)$ is itself an equivalence:

$$(A \simeq B) \simeq (A =_{\mathcal{U}} B). \tag{9}$$

**Definition 5.2** (Semantic Equivalence — Full Data)**.** Two semantic types $A$ and $B$ are *semantically equivalent* $(A \simeq B)$ if and only if there exist:

- Maps: $f : A \to B$, $g : B \to A$;
- Section: $\eta : \prod_{b:B} f(g(b)) =_B b$;
- Retraction: $\varepsilon : \prod_{a:A} g(f(a)) =_A a$;
- Coherence: $\tau : \prod_{a:A} \mathrm{ap}_f(\varepsilon(a)) =_{f(g(f(a)))=_B f(a)} \eta(f(a))$.

This is the full equivalence data. Cosine similarity captures only a shadow of this structure.

**Theorem 5.3** (Statistical Similarity $\neq$ Semantic Equivalence)**.** *Let* $\cos(F(A), F(B)) \geq 1 - \varepsilon$ *for embeddings* $F(A), F(B) \in \mathbb{R}^d$*. This does* ***not*** *entail the existence of maps* $f, g, \eta, \varepsilon, \tau$ *satisfying the full equivalence data of theorem 5.2.*

*Proof.* **The gap is structural, not quantitative.**

Cosine similarity is a *metric condition*: $d(F(A), F(B)) < \delta$, asserting proximity in a metric space. Semantic equivalence is an *algebraic condition*, requiring explicit maps and coherence data at all levels.

**Example (Frege's puzzle).** "The morning star" and "the evening star" have maximal embedding similarity (both reference Venus), so $\cos(F(\text{morning star}), F(\text{evening star})) \approx 1$. Yet the identity type carries non-trivial astronomical content—the identification requires the empirical discovery that the two celestial objects are the same planet. The coherence condition $\tau$ requires that the two "directions" of identification are compatible, which depends on the specific astronomical reasoning. This is invisible to any metric.

**Formal argument.** The space of embeddings within $\varepsilon$ of $F(A)$ is contractible (it is a ball in $\mathbb{R}^d$). But the space of semantic equivalences $A \simeq B$ is generically *non-contractible*: its $\pi_1$ corresponds to distinct auto-equivalences of $A$ (or $B$), which may be non-trivial. The canonical map from the contractible ball to the non-contractible equivalence space cannot be a homeomorphism, let alone an equivalence.

More precisely, the type of auto-equivalences $\mathrm{Aut}(A) \simeq (A \simeq A)$ has $\pi_0(\mathrm{Aut}(A)) = \mathrm{Out}(A)$ (outer automorphism group of $A$), which is non-trivial for semantically rich types. The ball of nearby embeddings has $\pi_0 = 0$. Hence no continuous map from the ball to the equivalence space can be surjective on $\pi_0$, and the metric condition cannot capture the algebraic structure. □

*Remark* 5.4 (The Univalence Constraint as Hallucination Filter). A model satisfying the univalence constraint would only identify two concepts $A$ and $B$ when the full equivalence data $(f, g, \eta, \varepsilon, \tau)$ exists—not merely when embeddings are close. This would eliminate:

- **Entity confusion:** "Morning star" $\neq$ "evening star" without explicit identification;
- **Attribute transfer:** Properties of $A$ are not assumed to transfer to $B$ without witnessing the appropriate transport map;
- **Analogical overextension:** Structural similarity (analogy) is not identity.

# 6 Persistent Homology for Practical Detection

The theoretical framework of the preceding sections establishes hallucination as a topological phenomenon. We now bridge to practical detection via persistent homology [3], which operates directly on the embedding spaces that LLMs produce.

**Definition 6.1** (Vietoris–Rips Complex)**.** Given a finite set of points $S \subset \mathbb{R}^d$ and a scale parameter $\varepsilon > 0$, the *Vietoris–Rips complex* at scale $\varepsilon$ is:

$$\mathrm{VR}_\varepsilon(S) = \{\sigma \subseteq S \mid d(v_i, v_j) \leq \varepsilon \text{ for all } v_i, v_j \in \sigma\}.$$

**Definition 6.2** (Persistence Diagram)**.** A *persistence bar* $(b, d, n)$ records the birth scale $b$, death scale $d$, and homological dimension $n$ of a topological feature in the Vietoris–Rips filtration $\{\mathrm{VR}_\varepsilon(S)\}_{\varepsilon \geq 0}$. The collection of all bars forms the *persistence diagram*.

The key insight is that persistent topological features (those with large persistence $d - b$) correspond to genuine structural holes in the knowledge base, not noise.

**Theorem 6.3** (Persistent Hallucination Detection)**.** *Let $\{v_t\}_{t=1}^T$ be the embedding sequence during text generation, and let $S_K$ be the set of knowledge base embeddings. Compute the Vietoris–Rips filtration on $\{v_t\} \cup S_K$. If there exists a persistent class $[\sigma] \in H_n$ with persistence $> \theta$ such that the generated path $\{v_t\}$ traverses the support of $[\sigma]$, then the generated text is hallucinating with confidence proportional to the persistence.*

*Proof.* **Step 1: Persistent features are genuine.** By the stability theorem of persistent homology (Edelsbrunner–Harer [3]), persistent features in the Vietoris–Rips filtration (those with persistence $\delta - \beta > \theta$ for a sufficiently large threshold $\theta$) correspond to genuine topological structure in the underlying space, not to noise-induced ephemeral features. Formally, for two point clouds $S, S'$ with Hausdorff distance $d_H(S, S') \leq \eta$, the bottleneck distance between their persistence diagrams satisfies:

$$d_B(\mathrm{dgm}(S), \mathrm{dgm}(S')) \leq \eta.$$

**Step 2: Persistent $H_1$ as semantic hole.** A persistent $H_1$ class corresponds to a stable loop in the knowledge embedding space that cannot be filled at any scale up to $\delta$ (the death scale). This is the embedding-space shadow of a non-trivial $H_1$ class in the knowledge complex $K_\bullet$ (theorem 2.13).

**Step 3: Traversal implies hallucination.** If the generation trajectory $\{v_t\}$ crosses this persistent loop—i.e., the sequence of embeddings traverses the support of $[\sigma]$—then the generated

text is navigating "empty semantic space" where no grounded knowledge provides support. Each step in the trajectory appears locally coherent (the embeddings are close to knowledge base embeddings), but the path as a whole traces a non-trivial cycle.

**Step 4: Confidence.** The confidence of the hallucination detection is proportional to the persistence $\delta - \beta$: longer-lived features represent more robust topological obstructions, and hence more confident hallucination detections. □

*Remark* 6.4 (Algorithmic Complexity). Computing the Vietoris–Rips filtration on $N$ points in $\mathbb{R}^d$ has worst-case complexity $O(2^N)$ due to the exponential growth of simplices. In practice, dimensionality reduction (e.g., via Johnson–Lindenstrauss projections) and lazy evaluation of the filtration (computing only simplices up to a fixed dimension $k$) reduce this to $O(N^{k+1})$. For $H_0$ and $H_1$ detection (the most common hallucination types), $k = 2$ suffices, giving $O(N^3)$ complexity—tractable for real-time monitoring of generation.

**Proposition 6.5** (Knowledge Extension Long Exact Sequence)**.** *Let $K_\bullet \hookrightarrow L_\bullet$ be an inclusion of knowledge complexes (adding new knowledge). The long exact sequence:*

$$\cdots \to H_n(K_\bullet) \to H_n(L_\bullet) \to H_n(L_\bullet, K_\bullet) \to H_{n-1}(K_\bullet) \to \cdots$$

*describes how adding knowledge can fill existing holes (killing elements of $H_n(K_\bullet)$) or create new holes (via relative homology $H_n(L_\bullet, K_\bullet)$).*

*Proof.* This is the standard long exact sequence of the pair $(L_\bullet, K_\bullet)$ in singular homology, applied to the simplicial knowledge complex. The connecting homomorphism $\partial_* : H_n(L_\bullet, K_\bullet) \to H_{n-1}(K_\bullet)$ identifies which new hallucination risks are introduced by knowledge extension: an element $[\tau] \in H_n(L_\bullet, K_\bullet)$ with $\partial_*([\tau]) \neq 0$ corresponds to a new knowledge element that, while filling one hole, opens another at a lower dimension. □

# 7 Haskell Formalization

We provide a Haskell formalization of the topological hallucination detection framework. The code listings below show the idealized type-theoretic structure; the companion executable module `TopologicalDetection.hs` uses a simplified string-based representation for computational tractability within Haskell's type system. The full GADT encoding with dependent type indices (as shown in the listings) would require a dependently-typed language such as Agda or Idris for complete static enforcement of the semantic invariants.

## 7.1 Semantic Types (Idealized)

The idealized semantic types mirror theorem 2.1:

Listing 1: Semantic types and terms (idealized GADT encoding).

```haskell
-- Idealized: in the companion code, SemTerm endpoints
-- are represented as String for Haskell compatibility.
data SemType where
```

```
  Entity   :: String → SemType
  Prop     :: String → SemType
  PathType :: SemType → SemTerm → SemTerm → SemType
  PathPath :: SemType → SemTerm → SemTerm → SemType
  ProdType :: SemType → SemType → SemType
  SumType  :: SemType → SemType → SemType
  VoidType :: SemType
  UnitType :: SemType

data SemTerm where
  Grounded     :: String → Source → SemTerm
  Inferred     :: SemTerm → Justification → SemTerm
  PathWitness  :: SemTerm → SemTerm → PathEvidence → SemTerm
  PathCompose  :: SemTerm → SemTerm → SemTerm
  Refl         :: SemTerm → SemTerm
  Hallucinated :: String → SemTerm
```

*Remark* 7.1 (Code Correspondence). The companion Haskell module `TopologicalDetection.hs` implements all five detectors and the type-checking judgment. Due to Haskell's lack of full dependent types, the implementation uses `String` identifiers where the idealized version uses `SemTerm` endpoints (e.g., `PathType SemType String String` rather than `PathType SemType SemTerm SemTerm`). The GADT and DataKinds extensions are declared for the algebraic data type definitions but do not enforce the full type-theoretic invariants at compile time. The semantic content and detection logic are faithful to the paper; only the static type-level guarantees are weaker.

## 7.2 The Five Topological Obstruction Detectors

Each detector implements the corresponding proof case from section 3:

Listing 2: Topological obstruction types (companion code uses String).

```
data TopologicalObstruction
  = MissingPath String String
    -- ^ pi_0 obstruction: no 1-cell connects these terms
  | NonContractibleLoop [String]
    -- ^ pi_1 obstruction: circular reasoning
  | IncoherentPaths String String PathEvidence PathEvidence
    -- ^ pi_2 obstruction: incompatible justifications
  | HomologicalHole Int [String]
    -- ^ H_n obstruction: n-cycle with no filling
  | TransportAnomaly String String String
    -- ^ Holonomy obstruction: compositional drift
```

## 7.3 Simplicial Knowledge Complex

Listing 3: Knowledge complex and homology.

```haskell
data KnowledgeComplex = KnowledgeComplex
  { kcVertices  :: [String]
  , kcEdges     :: [(Int, Int)]
  , kcTriangles :: [(Int, Int, Int)]
  }

-- Boundary operator: partial_1 : C_1 → C_0
boundary1 :: Chain1 → Chain0
boundary1 chain1 = Map.foldlWithKey' addBoundary Map.empty chain1
  where
    addBoundary acc (i, j) coeff =
      Map.insertWith (+) j coeff $
      Map.insertWith (+) i (-coeff) acc

-- Check if a 1-chain is a cycle (in ker partial_1)
isCycle1 :: Chain1 → Bool
isCycle1 chain = all (≡ 0) (Map.elems (boundary1 chain))
```

## 7.4 The Type-Checking Judgment

Listing 4: Type-checking as hallucination detection.

```haskell
typeCheck :: Context → SemTerm → SemType → Either
   HallucinationType Bool
typeCheck _ctx (Grounded claim _src) (Prop p)
  | claim ≡ p = Right True
  | otherwise  = Left (TypeMismatch claim p)
typeCheck _ctx (Inferred _base (Statistical prob)) _ty
  = Left (StatisticalNotProof prob)
typeCheck ctx (Inferred base (Entailment _ _)) ty
  = typeCheck ctx base ty
typeCheck _ctx (PathWitness a b NoEvidence) (PathType _ _ _)
  = Left (MissingPathDerivation a b)
typeCheck _ctx (Hallucinated _) _
  = Left FreeFloating
typeCheck _ _ _ = Left Unchecked
```

The full executable code, including a `main` function demonstrating all five hallucination types, is provided as the companion module `TopologicalDetection.hs`.

# 8 Discussion

## 8.1 Nature of the Result

The Hallucination–Homotopy Correspondence is *not* merely an analogy between topological concepts and LLM failures. It is a *functorial* relationship: the functor $\Phi : \mathbf{Hall} \to \mathbf{hTop}_*$ (theorem 3.13) provides a structure-preserving map from hallucination events to homotopy-theoretic invariants. This means:

1. Every hallucination has a topological *cause* (a non-trivial invariant);
2. Every non-trivial topological invariant in semantic space corresponds to a potential hallucination;
3. We *conjecture* that the classification is complete: the five cases of theorem 3.1 cover the homotopy-theoretic invariants that can obstruct faithful representation (see theorem 8.1).

**Conjecture 8.1** (Completeness of the HHC)**.** *The five cases of theorem 3.1—$\pi_0$, $\pi_1$, $\pi_2$, $H_n$, and holonomy—exhaust all hallucination types that arise from topological obstructions.*

*Remark* 8.2 (Evidence for Completeness). The evidence for theorem 8.1 rests on the Postnikov tower of the semantic $\infty$-groupoid $\mathbf{Sem}^{\mathrm{gpd}}$. Recall that any space $X$ admits a Postnikov tower:

$$\cdots \to X_{\langle 2 \rangle} \to X_{\langle 1 \rangle} \to X_{\langle 0 \rangle}$$

where $X_{\langle n \rangle}$ is the $n$-truncation with $\pi_k(X_{\langle n \rangle}) = 0$ for $k > n$. The homotopy type of $X$ is determined by the homotopy groups $\{\pi_n(X)\}_{n \geq 0}$ together with the $k$-invariants (Postnikov invariants) $k_n \in H^{n+2}(X_{\langle n \rangle}; \pi_{n+1}(X))$. Our five cases cover: $\pi_0$ (Case 2), $\pi_1$ (Case 1), $\pi_2$ (Case 3), the integral homology groups $H_n$ (Case 4, which subsumes higher homotopy groups via the Hurewicz theorem in the simply-connected case), and the fibration structure (Case 5, capturing coherence failures under transport). Higher homotopy groups $\pi_k$ for $k \geq 3$ correspond to higher-order coherence failures in justification chains; these are detected by the $H_n$ machinery of Case 4 via the Hurewicz homomorphism $h : \pi_n \to H_n$ (which is an isomorphism when $\pi_k = 0$ for $k < n$). A formal proof of completeness would require showing that the $k$-invariants introduce no hallucination types beyond those captured by the five cases. We leave this to future work.

## 8.2 The Categorical Attention Problem

Standard softmax attention computes:

$$\mathrm{Att}_{\mathbf{Vect}}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

which is a weighted colimit in **Vect** (a weighted sum). This always exists, even for contradictory inputs. The correct semantic attention would compute a weighted *limit* in **Sem**, which exists only when the diagram is *coherent* (the claims are mutually consistent). When the context contains $A$ and $\neg A$, no cone exists over the diagram, so the limit is empty. Generating output from contradictory context is producing a term of an empty type—hallucination by definition.

### 8.3 Sheaf-Theoretic Perspective

An alternative formulation uses sheaf cohomology. A generated claim $\hat{s}$ is hallucinated if and only if the obstruction class $o(\hat{s}) \in H^1(K_\bullet, \mathcal{F})$ is non-trivial, where $\mathcal{F}$ is the semantic sheaf over $K_\bullet$. This is dual to the homological formulation by the universal coefficient theorem: $H^1(K_\bullet, \mathcal{F}) \neq 0$ is dual to $H_1(K_\bullet) \neq 0$. Homology detects "holes" (missing justifications); cohomology detects "obstructions" (failures of local-to-global consistency).

### 8.4 Toward the HoTT-LM Architecture

The framework suggests five design principles for a hallucination-free language model:

1. **Generation as type inhabitation:** Replace $\arg\max_t P(t \mid \text{ctx})$ with the problem: find $a$ such that $\Gamma \vdash a : A$ has a derivation $\delta$.
2. **Certified derivations:** Every generated term $a : A$ comes with a derivation tree $\delta$ witnessing $\Gamma \vdash a : A$.
3. **Abstention on empty types:** When $A$ is uninhabited, the model returns $\perp$ rather than hallucinating.
4. **Context as fibration:** The context window is a dependent telescope, not a flat token sequence.
5. **Attention as limit:** The attention mechanism computes a weighted limit in **Sem** or signals incoherence, rather than silently blending contradictions.

### 8.5 Limitations and Open Problems

1. **Computability:** Type-checking in dependent type theory is undecidable in general. Restricting to decidable fragments (simple types, prenex quantification) trades expressiveness for tractability.
2. **Knowledge base completeness:** The framework assumes a well-defined knowledge complex $K_\bullet$. In practice, $K_\bullet$ is constructed from imperfect, incomplete knowledge bases. The persistent homology approach (section 6) mitigates this via stability.
3. **Computational cost:** Persistent homology computations scale polynomially but may be too slow for real-time token generation. Approximate methods (e.g., alpha complexes, witness complexes) offer trade-offs.
4. **Cubical type theory:** The cubical model of HoTT [6] provides a constructive interpretation of univalence, which may be more suitable for implementation than the classical formulation.
5. **The grounding problem:** Even with perfect topology, the initial grounding of $K_0$ (atomic facts) requires an oracle. Our framework assumes $K_0$ is given; constructing it reliably is an orthogonal challenge.

## 9 Conclusion

We have established a rigorous mathematical framework for understanding and detecting LLM hallucination through algebraic topology and homotopy type theory. The central insight is that

hallucination is not a statistical anomaly but a *structural inevitability* arising from the category-theoretic mismatch between the $\infty$-groupoid structure of semantic space and the contractible vector space representations used by current LLMs.

The five proof cases of the Hallucination–Homotopy Correspondence provide a classification—conjectured to be complete (theorem 8.1)—covering circular reasoning ($\pi_1$), unjustified inference ($\pi_0$), inconsistent justifications ($\pi_2$), fabricated entity chains ($H_n$), and compositional drift (holonomy). The Fundamental Obstruction Theorem shows that this is not a deficiency of particular architectures but of any system whose representation space is contractible. The Univalence Constraint formalizes why statistical similarity cannot substitute for semantic equivalence.

The persistent homology bridge provides a practical path from theory to detection: topological features in embedding space, when persistent, signal genuine structural holes corresponding to hallucination-prone regions. The Haskell formalization demonstrates that these concepts are not merely abstract but computationally tractable.

The path toward hallucination-free language models passes through richer representation spaces—spaces that faithfully preserve the homotopy-theoretic structure of meaning. Whether this takes the form of type-theoretic generation, topological monitoring, or hybrid architectures, the mathematics is clear: *the topology of meaning must be respected, not collapsed.*

## References


[1] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, 36(1–4):345–384, 2010.

[2] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study, 2013.

[3] H. Edelsbrunner and J. L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[4] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232*, 2023.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[6] C. Cohen, T. Coquand, S. Huber, and A. Mörtberg. Cubical Type Theory: A Constructive Interpretation of the Univalence Axiom. *Journal of Automated Reasoning*, 60(2):159–194, 2018.

[7] J. Lurie. *Higher Topos Theory*. Princeton University Press, 2009.

[8] J. P. May. *A Concise Course in Algebraic Topology*. University of Chicago Press, 1999.

[9] E. Riehl. *Categorical Homotopy Theory*. Cambridge University Press, 2014.

[10] G. Carlsson. Topology and Data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[11] R. Ghrist. Barcodes: The Persistent Topology of Data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[12] S. Awodey and M. A. Warren. Homotopy Theoretic Models of Identity Types. *Mathematical Proceedings of the Cambridge Philosophical Society*, 146(1):45–55, 2009.

[13] M. Shulman. Univalence for Inverse Diagrams and Homotopy Canonicity. *Mathematical Structures in Computer Science*, 25(5):1203–1277, 2015.

[14] A. Tonks. Relating the Associahedron and the Permutohedron. In *Operads: Proceedings of Renaissance Conferences*, pages 33–36, 1997.

[15] G. Friedman. *Survey Article: An Elementary Illustrated Introduction to Simplicial Sets. Rocky Mountain Journal of Mathematics*, 42(2):353–424, 2012.