A Geometric Taxonomy of Hallucination in Large Language Models

Anonymous ACL submission

Abstract

The term "hallucination" is used to describe at least three different failure modes in LLMs: ignoring provided context, inventing semantically foreign content, and producing factually wrong details within a plausible conceptual frame. These differences are not merely taxonomic—they produce different geometric signatures in embedding space, and that geometry determines what is detectable and what is not.

We propose a three-type taxonomy grounded in this observation, and introduce two detection methods derived from it. The Semantic Grounding Index (SGI) detects semantic laziness—responses that remain near the query instead of moving toward source material on the unit hypersphere. The Directional Grounding Index (DGI) detects confabulation by measuring whether response displacement aligns with the typical grounding direction learned from verified examples.

Both require a single embedding call and no secondary language model. Experiments across five benchmarks confirm the taxonomy's predictions: semantic laziness and confabulation are geometrically detectable; factual errors within correct conceptual frames are not, for principled reasons rooted in how distributional representations encode meaning. Apparent classifier success on factual-error data is traced to an annotation style confound, not genuine geometric signal.

Detection is domain-local: a grounding direction learned in one domain does not transfer to another unless the underlying error type is consistent.

1 Introduction

LLMs generate text by optimizing distributional statistics without maintaining correspondence to external reality (Brown et al., 2020; Achiam et al., 2023). This produces hallucination—fluent, confident text that is factually wrong (Ji et al., 2023; Huang et al., 2023)—posing challenges for deployment in high-stakes contexts where accuracy must be documented (European Parliament and Council of the European Union, 2024).

"Hallucination" refers to at least three different failure modes. A model that ignores provided context (Maynez et al., 2020) differs fundamentally from one that invents non-existent entities, which differs from one that provides wrong details about real concepts (Zhang et al., 2023). These pathologies have different roots, different consequences, and different geometric signatures in embedding space.

The problem is compounded by how hallucination benchmarks are built. Most evaluation datasets generate false answers by prompting LLMs to fabricate (Li et al., 2023), which captures how models write when asked to lie rather than the geometry of genuine factual error. Hicks et al. (2024) have questioned whether "hallucination" is even the right framing, proposing instead "bullshit" in Frankfurt's sense (Frankfurt, 2005)—speech produced without concern for truth. Rather than settling this terminological debate, we operationalize differences geometrically: if different failure modes produce different signatures on $\mathbb{S}^{d-1}$, the taxonomy becomes empirically grounded.

We introduce SGI (Semantic Grounding Index) for context-based detection of Type I, and DGI (Directional Grounding Index) for context-free detection of Type II. Both are grounded in a formal analysis of what contrastive training implies about the geometry of grounded and ungrounded responses (Section 3.1). Through experiments on five benchmarks, we characterize where geometric detection succeeds, where it fails, and why. The three types are defined as follows:

Type I (Unfaithfulness) The model ignores provided context, generating from parametric memory (Maynez et al., 2020). The response remains angu-

larly close to the query rather than moving toward the context. We term this *semantic laziness*: the model defaults to question-proximate completions rather than venturing into the context's semantic territory. Detectable via SGI (Section 3.2).

Type II (Confabulation) The model invents entities, mechanisms, or concepts that do not exist. The response displacement deviates from the plausible-answer region of $\mathbb{S}^{d-1}$ in geometrically anomalous directions. Detectable via Γ (Section 3.2).

Type III (Factual Error) The model provides wrong details within the correct conceptual frame. Because distributional representations encode co-occurrence, not truth conditions (Harris, 1954; Firth, 1957), a factually wrong response about a topic occupies the same region of $\mathbb{S}^{d-1}$ as a correct one. Geometrically undetectable by construction (Section 4.3).

2 Related Work

The word "hallucination" entered NLP from neural machine translation, where it described outputs unrelated to source sentences (Lee et al., 2018; Raunak et al., 2021). Academic definitions cluster around two poles: Ji et al. (2023) emphasize unfaithfulness to source; Huang et al. (2023) emphasize plausible-yet-nonfactual content. Maynez et al. (2020) identifying intrinsic hallucinations (contradicting the source) from extrinsic ones (unverifiable from source). Our taxonomy refines this space: Type I corresponds roughly to intrinsic hallucination; Type II to a specific class of extrinsic hallucination where semantic displacement is large; Type III to factual error within a plausible semantic frame, which neither prior taxonomy formally categorize.

Approaches requiring model internals achieve strong performance: hidden-state trajectories identify truthful from hallucinated generations (Azaria and Mitchell, 2023) (AUROC 0.96); eigenvalue spectra of response covariance provide complementary signals (Chen et al., 2024). These require white-box access unavailable in API deployments. Consistency-based methods sample repeatedly: SelfCheckGPT (Manakul et al., 2023) achieves AUC-PR 0.93; semantic entropy (Farquhar et al., 2024; Kuhn et al., 2023) achieves AUROC 0.79, both requiring 5–20 generations per query. When source documents are available, alignment verification works well: HHEM (Vectara, 2024) achieves 74–77% balanced accuracy; AlignScore (Zha et al., 2023) reaches AUC-ROC 0.87. Our proposed methods require neither model internals, nor multiple generations, nor source documents—using only the geometry of a single embedding call.

Contrastive training explicitly optimizes angular relationships on $\mathbb{S}^{d-1}$ (Reimers and Gurevych, 2019; Gao et al., 2021). Wang and Isola (2020) analyzes alignment and uniformity as complementary objectives of contrastive learning: the alignment term pulls semantically related pairs together on the hypersphere, while the uniformity term distributes representations across its surface. Our approach exploits the directional consequence of this structure: when a response is grounded in context, the alignment objective produces a systematic displacement from the query toward the answer region. Ungrounded responses do not undergo this directed movement, which is what SGI and DGI measure.

3 Methods

3.1 Geometric Motivation

We motivate SGI and DGI by analyzing what contrastive training implies about the geometry of grounded responses on $\mathbb{S}^{d-1}$.

The alignment objective of contrastive training pulls semantically similar pairs $(\phi(q), \phi(r))$ close together on the hypersphere (Wang and Isola, 2020). When a response r is genuinely grounded in context c, it draws content from c and therefore becomes semantically aligned with c. The angular trajectory from q to r—the displacement $\delta(q, r) = \hat{\phi}(r) - \hat{\phi}(q)$—points toward the region occupied by c. A semantically lazy response that ignores c and generates from parametric memory remains near q, producing small or null displacement toward c. This is the geometric basis for SGI.

For context-free applications, the argument is distributional. When a model correctly answers a question q, the embedding of r reflects the semantic neighborhood of q's expected answer. Across many grounded (q, r) pairs, the displacements $\hat{\delta}(q_i, r_i)$ cluster directionally—each response moves away from the query in a domain-consistent direction. The Von Mises–Fisher mean of these directions defines the typical grounding trajectory $\hat{\mu}$ (Banerjee et al., 2005). A confabulated response invents content that is semantically foreign to that trajectory, producing displacement with low or negative alignment to $\hat{\mu}$. This is the geometric basis for Γ.

Type III factual errors are geometrically invisible for a principled reason. The distributional hypothesis (Harris, 1954; Firth, 1957) holds that semantic representations encode co-occurrence statistics, not propositional content. A response claiming "the capital of Australia is Sydney" occupies the same semantic neighborhood as "the capital of Australia is Canberra"—both involve Sydney, Canberra, capitals, and Australia in similar distributional contexts. No angular operation can distinguish them, because the embedding model was not trained to encode truth conditions.

Proposition 3.1. Under any representation model trained on the distributional hypothesis, factual errors of Type III (within-frame substitutions) are not separable from correct responses in $\mathbb{S}^{d-1}$ by angular geometry alone.

We treat Proposition 3.1 as a working hypothesis rather than a formal theorem; the TruthfulQA experiments in Section 4.3 provide empirical support, and Section 5 discusses the theoretical argument further.

3.2 Detection Methods

Let $\hat{\phi}(s) = \phi(s)/\|\phi(s)\|$ denote the L_2-normalized sentence representation. Angular distance $\theta(a, b) = \arccos(\hat{\phi}(a)^\top \hat{\phi}(b))$ is the geodesic distance on $\mathbb{S}^{d-1}$ (do Carmo, 1992; Bridson and Häfliger, 1999) and satisfies the triangle inequality, which constrains how the three pairwise distances $\theta(q, c)$, $\theta(q, r)$, $\theta(r, c)$ can simultaneously take values.

Semantic Grounding Index (SGI) For query q, context c, and response r:

$$\mathrm{SGI}(r;\, q, c) = \frac{\theta(r, q)}{\theta(r, c)} \tag{1}$$

A grounded response moves toward the context's region on $\mathbb{S}^{d-1}$ (Section 3.1), satisfying SGI > 1. A semantically lazy response remains near the query, satisfying SGI ≤ 1. The ratio compares two geodesic distances on the same sphere and is not bounded by $[0, 1]$. Angular distance is preferred over cosine similarity because the geodesic metric satisfies the triangle inequality (Bridson and Häfliger, 1999), making the ratio geometrically principled.

Directional Grounding Index (Γ) For query q and response r, the normalized displacement is

$$\hat{\delta}(q, r) = \delta/\|\delta\| \tag{2}$$

where $\delta = \hat{\phi}(r) - \hat{\phi}(q)$. Given a reference set $\mathcal{R} = \{(q_i, r_i)\}_{i=1}^N$ of verified grounded pairs:

$$\hat{\mu} = \mu/\|\mu\| \tag{3}$$

where $\mu = \frac{1}{N}\sum_{i=1}^N \hat{\delta}(q_i, r_i)$. Normalizing each $\hat{\delta}$ before averaging computes the Von Mises–Fisher mean direction on the sphere (Banerjee et al., 2005), ensuring $\hat{\mu}$ reflects the typical grounding direction rather than displacement magnitude. The two normalizations serve distinct geometric purposes and cannot be collapsed.

We define the Directional Grounding Index as

$$\Gamma(q, r;\, \mathcal{R}) = \hat{\delta}(q, r)^\top \hat{\mu} \tag{4}$$

with $\Gamma \in [-1, +1]$. High values indicate alignment with the learned grounding direction; low or negative values indicate displacement anomalous to that direction, characteristic of Type II confabulation. A local variant uses k-nearest neighbors in $\mathcal{R}$ to compute a query-adaptive $\hat{\mu}_q$. Global Γ requires one embedding and one dot product: $O(d)$.

3.3 Baseline and Statistics

We compare Γ against an NLI baseline in all reference-free experiments (Type II and Type III). NLI is the strongest available black-box alternative that operates on the same inputs as Γ: it requires only a (q, r) pair with no model internals, no multiple generations, and no source document. We use `cross-encoder/nli-deberta-v3-small`, which outputs an entailment probability that serves as a grounding score. SGI is evaluated without a baseline: it operates on (q, c, r) triples, and no standard entailment model accepts this three-input configuration. The NLI comparison is therefore scoped to settings where both methods receive identical inputs, ensuring a fair evaluation.

We report AUROC with 1,000-resample bootstrap CIs, permutation test p-values (5,000 resamples), and Hanley–McNeil z-tests (Hanley and McNeil, 1982). All primary experiments use `sentence-t5-large` (768 dimensions); architecture ablation appears in Section 4.2.4.

4 Results

4.1 Type I: SGI on Grounded Context

Table 1 reports SGI on HaluEval QA ($n = 10{,}000$) (Li et al., 2023) across five embedding architectures. Grounded responses satisfy SGI > 1; semantically lazy (Type I) responses satisfy SGI ≤ 1.

Cross-model correlation ($r = 0.85$) confirms SGI measures a property of the text, not an embedding artifact. Mean SGI for grounded responses (1.180) are higher than 1 across all models, confirming the Section 3.1 prediction that grounded responses move toward context on $\mathbb{S}^{d-1}$. Semantically lazy responses (mean 0.910) remain near the query—the signature predicted by the semantic laziness hypothesis.

Table 1: SGI on HaluEval QA (n=10,000). Grounded SGI > 1; semantically lazy (Type I) SGI ≤ 1. AUROC 0.776–0.824 consistent across architectures (r=0.85 cross-model).

Model	Grnd.	Halluc.	AUROC	Recall@5%
sentence-t5-large	1.203	0.856	0.824	0.41
all-mpnet-base-v2	1.142	0.921	0.776	0.33
bge-base-en-v1.5	1.231	0.948	0.823	0.40
e5-base-v2	1.138	0.912	0.794	0.37
all-MiniLM-L6-v2	1.188	0.913	0.806	0.38
Mean	1.180	0.910	0.805	0.38

4.2 Type II: DGI on Confabulations and External Benchmarks

4.2.1 Human-crafted confabulations

Table 2 reports Γ on 142 human-written confabulations (finance, medical, legal) designed to represent prototypical Type II errors: redefined terms, fabricated institutions, invented mechanisms. These examples introduce semantically foreign content in the sense of Section 3.1—displacement off the plausible-answer region of $\mathbb{S}^{d-1}$—unlike within-frame substitutions (Type III).

Table 2: Γ on human-crafted confabulations (n=142) vs. NLI. Cross-domain degradation 3.8%. Local/global difference 0.7% (n.s.).

Method	AUROC	95% CI	Recall@10
Γ global $\hat{\mu}$	0.958	[0.924, 0.992]	0.84
Γ local (k=15)	0.965	[0.931, 0.999]	0.87
NLI CrossEncoder	0.611	[0.527, 0.695]	0.19

Γ outperforms NLI by $\Delta = 0.347$. NLI achieves AUROC=0.611 because confabulations are syntactically coherent and entailment-compatible (Zha et al., 2023). Surface entailment cannot detect anomalous displacement off the plausible-answer region of $\mathbb{S}^{d-1}$.

4.2.2 Domain-locality on LLM-generated benchmarks

The domain-locality finding requires a theoretical context before the numbers are interpretable. Contrastive training on domain-specific corpora induces domain-specific grounding directions $\hat{\mu}$. There is no guarantee that $\hat{\mu}_{\text{finance}}$ and $\hat{\mu}_{\text{dialogue}}$ are aligned—they reflect the distributional geometry of different vocabularies and context patterns. Detection therefore works within domains but not across them.

On HaluEval QA and Dialogue (Li et al., 2023), in-domain AUROC is 0.68–0.99; cross-domain AUROC decreases to $\approx$0.50 (Dialogue $\rightarrow$ QA: 0.38). The mean cosine similarity between domain grounding directions is $\bar{\cos}(\hat{\mu}) = -0.07$, confirming near-orthogonality. Human-crafted confabulations show only 3.8% degradation because real semantic departure from the plausible-answer region is domain-agnostic: an invented mechanism is anomalous regardless of the domain. Prompted fabrication (Li et al., 2023) produces domain-specific stylistic artifacts that do not transfer.

4.2.3 External human-annotated benchmarks

The domain-locality finding raises a question: does real confabulation share a consistent geometry across domains? Table 3 provides evidence. Three independently collected datasets were used, none involved in method development. WikiBio GPT-3 (Manakul et al., 2023) provides 238 GPT-3 biographies with sentence-level human annotation (102 matched pairs). FELM (Chen et al., 2023) provides segment-level factuality labels across world-knowledge and writing domains (81 pairs). ExpertQA (Malaviya et al., 2024) provides claim-level expert annotations across 32 specialist fields (900 pairs).

In ExpertQA experiment Γ outperforms NLI by $\Delta = 0.243$ on expert-annotated inaccurate claims across 32 specialist fields. NLI is at chance (0.452, $p = 1.0$) because expert errors are entailment-compatible but occupy a different region of embedding space. This confirms that Γ detects a displacement that surface entailment cannot see, and that the advantage scales with the specificity of the domain.

FELM experiment shows similar modest domain-specific detection (0.648, $p = .013$) but slightly higher than NLI result (0.624). WikiBio domain-specific detection fails (0.581, $p = .199$) being NLI a stronger method. The explanation con-

Table 3: External benchmark validation: domain-specific Γ vs. NLI CrossEncoder. p: permutation test ($p<.05$ in bold). Δ: $\Gamma_{\text{dom}} - \text{NLI}$.

Dataset	Γ dom.	95% CI	p	NLI	95% CI	Δ
WikiBio (n=102)	0.581	[0.416, 0.695]	.199	**0.712**	[0.639, 0.781]	−0.131
FELM (n=81)	**0.648**	[0.510, 0.801]	**.013**	0.624	[0.532, 0.704]	+0.024
ExpertQA (n=900)	**0.695**	[0.654, 0.738]	**.000**	0.452	[0.426, 0.478]	+0.243

nects directly to the taxonomy: WikiBio's annotation criterion marks any incorrect detail as "major inaccurate" regardless of semantic distance, conflating Type II errors (invented content with large displacement) and Type III errors (within-frame substitutions with near-zero displacement). When a dataset conflates the two types, Γ detects only the subset that are genuine Type II.

Table 4 reports the cross-domain transfer matrix. FELM↔ExpertQA symmetry (0.822 in both directions) confirms that open-ended factual confabulation shares a consistent geometric regime across subject domains. WikiBio's $\hat{\mu}$ actively misleads on both other datasets (AUROC 0.200 and 0.288, well below chance), identifying it as geometrically incommensurable with the confabulation regime. This asymmetry provides the strongest evidence that WikiBio samples a different geometric population than genuine Type II confabulation.

Table 4: Cross-domain transfer (AUROC). FELM↔ExpertQA symmetry (0.822) confirms a shared geometric regime for factual confabulation. WikiBio's $\hat{\mu}$ misleads on both other datasets.

	Test domain		
Train $\hat{\mu}$	WikiBio	FELM	ExpertQA
WikiBio	0.537	0.200	0.288
FELM	0.433	**0.822**	0.653
ExpertQA	0.446	0.822	**0.691**

4.2.4 Architecture ablation

An important question for any embedding-based method is whether it measures a property of the text or an artifact of a particular model's geometry. We evaluate Γ across three sentence transformers that vary in both dimensionality and training procedure (Table 5).

Two effects are separable. First, dimensionality: all-MiniLM-L6-v2 (384d, AUROC 0.800) is outperformed by both 768-dimensional models, consistent with higher-dimensional spaces providing finer resolution for displacement measurement. Second, architecture: sentence-t5-large and all-mpnet-base-v2 share the same dimensionality yet differ by 0.078 AUROC (0.958 vs. 0.880). This gap is larger than the dimensionality gap, indicating that encoder depth and contrastive training shape the sharpness of the learned grounding direction more than dimension alone.

If the signal were an artifact of any one model's geometry, scores would diverge across architecturally distinct models. Instead, all three models agree in direction: grounded pairs produce higher Γ than confabulations in every case, with non-overlapping 95% CIs between the best and worst performers. As we show in concurrent work, these results mirror the cross-model robustness established for SGI, where five architecturally different models achieved mean pairwise Pearson correlation $r = 0.85$. Convergence across independently trained models with different architectures on different objectives is not expected unless the displacement geometry reflects something intrinsic to the text.

Table 5: Architecture ablation on human-crafted confabulations (n=142). Dimension and training procedure both contribute independently to AUROC; non-overlapping CIs confirm the sentence-t5-large advantage is reliable.

Model	Dim.	AUROC	95% CI
all-MiniLM-L6-v2	384	0.800	[0.730, 0.870]
all-mpnet-base-v2	768	0.880	[0.820, 0.940]
sentence-t5-large	768	0.958	[0.924, 0.992]

4.3 Type III Boundary: TruthfulQA

TruthfulQA (Lin et al., 2022) provides 817 questions with human-written truthful (`best_answer`) and false (`incorrect_answers`) response pairs. False answers reflect beliefs people actually hold and are topically identical to truthful ones—the defining characteristic of Type III: wrong detail with correct conceptual frame. Together with DGI and NLI we use a Logistic Regression classifier (LR) as a basic baseline. Table 6 reports all these methods.

The LR achieves AUROC 0.731, which appears

Table 6: Detection on TruthfulQA (817 matched pairs). Bootstrap 95% CI. HM: Hanley–McNeil z-test (Hanley and McNeil, 1982). Cosine inversion (AUROC 0.365) is the diagnostic finding; LR signal is a style artifact.

Method	**AUROC**	**95% CI**	p_{perm}	p_{HM}	**Signal**
Cosine similarity	0.365	[0.340, 0.392]	1.000	$< .001$	inverted
Embedding magnitude	0.518	[0.497, 0.541]	.059	.198	not significant
Γ (80/20 split)	0.535	[0.473, 0.600]	.137	.265	not significant
Γ (global $\hat{\mu}$)	0.579	[0.551, 0.607]	.000	$< .001$	style
NLI entailment	0.539	[0.511, 0.567]	.004	.006	style
LR raw embeddings (CV)	0.731	[0.706, 0.756]	.000	$< .001$	style artifact
LR on Γ features	0.786	[0.765, 0.807]	.000	$< .001$	style artifact

to contradict the Type III hypothesis. The cosine inversion resolves this: false answers are geometrically closer to queries than truthful ones (AUROC 0.365, below chance)—the opposite of confabulation geometry. Proposition 3.1 predicts no displacement signal. The inversion goes further, suggesting an active anti-correlation between falseness and angular distance. The explanation could be annotation style: truthful answers in TruthfulQA are longer and hedged, and false answers are shorter and declarative. Longer responses accumulate semantic content orthogonal to the query, producing larger displacement vectors that the LR exploits across 768 dimensions.

Held-out Γ calibration, which cannot propagate this style signal through $\hat{\mu}$, yields AUROC 0.535 ($p = .265$, CI contains 0.50). The LR detects how TruthfulQA was written, not whether answers are factually correct. Three observations confirm this interpretation: (1) the cosine signal is inverted relative to what factual-error detection would predict; (2) held-out Γ yields AUROC 0.535 ($p = .265$); (3) per-category Γ averages 0.823 but inflates with small-sample overfitting—categories with $n > 50$ converge toward 0.57–0.65, consistent with a modest style signal rather than factual detection.

5 Discussion

Taxonomy validated by geometry The experiments confirm that the three types of hallucination produce distinct geometric signatures that determine detectability. Type I unfaithfulness produces a measurable failure to move toward context on $\mathbb{S}^{d-1}$—semantic laziness—captured by SGI with AUROC 0.805 across five architectures. Type II confabulation produces anomalous displacement from the plausible-answer region, captured by Γ with AUROC 0.958 on human-crafted instances and $\Delta = 0.243$ over NLI on expert-domain data. Type III factual error produces no geometric anomaly, because a wrong detail within the correct conceptual frame occupies the same region of $\mathbb{S}^{d-1}$ as the correct detail. The taxonomy does not merely name these differences—it predicts them from the geometry of how embedding models represent meaning.

Type III as a theoretical constraint The TruthfulQA analysis establishes that the apparent detection signal (LR AUROC 0.731) is geometrically inconsistent with factual-error detection. False answers are closer to queries than truthful ones, the opposite of what any displacement-based method would produce under genuine error detection. The held-out Γ result (AUROC 0.535, $p = .265$) confirms that once annotation style is removed, geometry offers nothing. This follows from Proposition 3.1 and the distributional hypothesis (Harris, 1954; Firth, 1957): embedding geometry encodes co-occurrence, not truth conditions. Type III errors are geometrically invisible by construction, not by methodological failure.

Domain-locality and its limits The near-orthogonality of domain grounding directions ($\bar{\cos}(\hat{\mu}) = -0.07$) explains the collapse of cross-domain detection on LLM-generated benchmarks. Each domain's contrastive training history induces a different grounding direction; there is no domain-agnostic $\hat{\mu}$ that captures all of them simultaneously. The FELM↔ExpertQA transfer symmetry (0.822) is therefore a finding, not an assumption: it shows that open-ended factual confabulation—as annotated by domain experts—shares a geometric regime that LLM-generation artifacts do not. The WikiBio result, where $\hat{\mu}$ trained on biography text actively misleads on both FELM and ExpertQA (AUROC 0.200 and 0.288), identifies the annotation criterion, not the domain, as the relevant variable.

Practical implications Γ with domain-specific calibration provides first-line confabulation detec-

tion where NLI fails, at $O(d)$ cost per query. The ExpertQA result ($\Delta = 0.243$) is directly actionable for legal, medical, and financial applications, where expert-domain inaccuracies are precisely the failure mode that entailment-based methods cannot see. For RAG systems with source documents, SGI addresses the complementary case: when context is available, semantic laziness is directly measurable, and a Type III error that is undetectable in isolation becomes a Type I context-contradiction if the correct answer appears in the source. For AI compliance documentation (European Parliament and Council of the European Union, 2024), the taxonomy provides vocabulary for precise, auditable capability claims: a system validated for confabulation detection cannot claim to detect all hallucination types, and the geometric argument explains why.

6 Conclusion

We introduce a geometric taxonomy of LLM hallucination identifying three failure modes by their signatures on $\mathbb{S}^{d-1}$, and present two detection methods grounded in this taxonomy. SGI detects Type I semantic laziness by measuring whether a response moves toward provided context. Γ detects Type II confabulation via displacement geometry, achieving AUROC 0.958 ± 0.034 on human-crafted confabulations and 0.581–0.695 on external human-annotated benchmarks (Manakul et al., 2023; Chen et al., 2023; Malaviya et al., 2024), outperforming NLI by $\Delta = 0.243$ on expert-domain data where surface entailment operates at chance. Type III factual errors resist geometric detection; apparent classifier success on TruthfulQA (Lin et al., 2022) (LR AUROC 0.731) is traced to a stylistic annotation confound—false answers are geometrically closer to queries than truthful ones (cosine 0.365, inverted)—confirming a theoretical constraint rather than a methodological gap. The domain-locality finding ($\bar{\cos}(\hat{\mu}) = -0.07$) and FELM↔ExpertQA transfer symmetry (0.822) characterize when geometric detection generalizes: within genuine confabulation regimes, a consistent grounding direction exists and transfers; within LLM-generation- artifact regimes, it does not. The taxonomy predicts which side of this boundary any dataset occupies.

Reproducibility All embedding and NLI models are publicly available on HuggingFace (Reimers and Gurevych, 2019). Benchmarks HaluEval (Li et al., 2023), TruthfulQA (Lin et al., 2022), WikiBio GPT-3 (Manakul et al., 2023), FELM (Chen et al., 2023), and ExpertQA (Malaviya et al., 2024) are publicly accessible. The human-crafted confabulation dataset and code will be released upon publication. Hyperparameters: $k = 15$, bootstrap $B = 1{,}000$, permutation $M = 5{,}000$.

Limitations

The human-crafted confabulation dataset ($n = 142$) prepared for this research covers prototypical Type II errors with clear semantic departure. Mixed confabulations—responses that partially engage with context while introducing fabricated elements—may produce intermediate Γ scores whose interpretation is not yet characterized. The dataset covers three domains (finance, medical, legal); generalization to code generation, conversational, or multilingual confabulation remains untested.

The Type II/III boundary is defined geometrically but not yet formally proven. Proposition 3.1 is stated as a working hypothesis supported by the TruthfulQA evidence; a formal proof on $\mathbb{S}^{d-1}$ for within-frame substitutions—independent of the distributional hypothesis—remains open.

The taxonomy covers outputs of generative models. It does not address retrieval errors, prompt injection, or failures that occur before generation. The experiments use English-language benchmarks; whether the geometric signatures of grounding and confabulation transfer across languages is an open question.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

Martin R Bridson and André Häfliger. 1999. *Metric spaces of non-positive curvature*, volume 319. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. FELM: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems*.

Manfredo Perdigão do Carmo. 1992. *Riemannian Geometry*. Birkhäuser.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*, L.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic consistency. *Nature*, 630:625–630.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*, pages 1–32.

Harry G Frankfurt. 2005. *On bullshit*. Princeton University Press.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2–3):146–162.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology*, 26:38.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Grangier. 2018. Hallucinations in neural machine translation. In *NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, Dan Roth, and Chris Clark. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2021*, pages 1172–1183.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Vectara. 2024. HHEM: Hughes hallucination evaluation model. *HuggingFace Model Repository*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9929–9939.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11328–11348.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.