Hallucinations in Large Language Models:

When More Data Reinforces Error and Minimal Perturbations Reveal Misclassification

Structured Abstract

Problem Definition: Large language models continue to exhibit low-accuracy outputs that are conventionally labeled “hallucinations.” In high-liability domains (finance, legal reasoning, clinical decision support, regulatory compliance) even isolated fabrications carry severe material risk. Aggregate hallucination-rate metrics appear to improve under scaling, alignment, and retrieval augmentation, yet practitioners encounter persistent surprises that evade these interventions. The core issue is that “hallucination” functions as a heterogeneous catchall term that conflates causally distinct error classes. Current evaluation and mitigation frameworks treat the phenomenon as a single scalar, obscuring the fact that most observed errors are resolvable through minimal prompt clarification while a narrow irreducible tail persists under exhaustive specification. This misclassification blocks reliable deployment in zero-tolerance environments and prevents the technology from reaching economically significant use cases.

Proposed Contribution: This paper introduces a prompt-level diagnostic fork that separates resolvable trajectory errors (arising from underspecification, corpus skew, or compression artifacts) from true hallucinations (narrow, irreducible fabrications that remain incoherent under any plausible binding). The contribution is a zero-cost, model-agnostic perturbation protocol together with a conservative gate criterion and an orthogonal severity matrix. These primitives enable precise triage: most errors can be resolved at the surface level, while the irreducible tail is graded by practical consequence rather than abstract correctness. The framework is reductionist by design — it simplifies the problem from an intractable scalar to a set of separable, measurable classes.

Theoretical Foundations: Trajectory errors are statistically coherent generations that follow a high-probability lexical path mismatched to the speaker’s intended frame. True hallucinations are generations whose conditional probability approaches zero under any binding of the prompt variables, implying an orthogonal internal disruption outside the learned manifold. The fork is operationalized via a gate requiring adequate specification and evidence conflict. Severity is graded on four continuous axes (Evidence Conflict, Materiality/Risk, Detectability/Recoverability, with Specification Adequacy recorded but excluded from the scalar index) and mapped to fuzzy tolerance ranges that support abstention, verification, or logging decisions.

Cross-Domain Mapping: constraint topology — alignment dynamics — uncertainty modeling — structural inference — macro-to-micro mapping — probabilistic cognition — trajectory formation — semiotic frustration — specification adequacy — severity triage — error-tracking taxonomy — diagnostic primitives

Scope and Intent: The paper provides a conceptual and operational foundation for diagnosing and prioritizing hallucinations. It deliberately omits deeper tensor-logic analysis (scar accumulation, edge fragility, parallelism-induced decoherence) that appears in related work, as that material has not yet been cleared for public release. The intent is to offer immediately usable primitives for practitioners while establishing a scaffold for future integration with structural models. No experiments, no architectural proposals, no claim of completeness — only a cleaner way to ask the right question.

ORCID: 0009-0004-5339-5373

DOI: 10.5281/zenodo.18674477

Keywords:

hallucination fork, trajectory error, true hallucination, semiotic frustration, prompt perturbation, specification gate, severity matrix, tolerance range, sense-binding rate, diagnostic primitive, error-tracking taxonomy, irreducible tail, comprehension limit, liability threshold

Orientation for Interpretation:

This paper proposes conceptual primitives and an operational diagnostic framework. It is not clinical validation, empirical benchmark, or architectural proposal. The primitives are reductionist by design — intended to simplify classification so that future empirical, modeling, and cross-domain work can build on a clearer foundation. Claims are provisional and scoped to prompt-level observability. Terminology is precise yet domain-general. Readers should expect abstraction before application and diagnostic clarity rather than complete mechanistic explanation. The deeper tensor-level account of the irreducible tail is preserved for qualified partners.

Section 1: Introduction

Low-accuracy generations in large language models remain one of the most persistent and costly problems in deployment. In high-stakes domains — clinical decision support, legal research, financial analysis, regulatory compliance, customer-facing chat — even a single incorrect or fabricated assertion can lead to material harm, eroded trust, or regulatory exposure. Benchmarks report aggregate "hallucination rates" that appear to decline with scale, alignment, or retrieval, yet practitioners continue to encounter surprises that evade those same improvements. The gap between reported progress and real-world experience suggests that current measurement and mitigation approaches are addressing only part of the phenomenon.

In this paper we use a conservative working notion: *hallucination* refers to low-accuracy outputs that persist under adequate specification and contradict accessible evidence. The exhibits that follow motivate this distinction.

As a result "hallucination" has become a catchall label for any factually incorrect, implausible, or contextually mismatched output. Evaluation suites treat it as a single scalar and progress is measured by reductions in that scalar. Yet when the same low-accuracy output is probed with small, targeted changes to the input prompt, it often flips to high accuracy under one edit and remains unchanged under others. This, in our opinion, is evidence that the bucket contains multiple, distinguishable error classes.

The core question that has received surprisingly little explicit attention is therefore:

"What exactly distinguishes a wrong answer from a hallucination"

In ordinary language use, humans routinely produce factually incorrect statements that are nonetheless coherent and grammatically valid — these are miscommunications that arise when speaker and listener follow different implicit assumptions or resolve a shared term in different ways. The same speaker can correct the mismatch with a single clarifying sentence. In contrast, clinical psychology defines hallucination as the perception of something that has no external correlate at all — an experience disconnected from shared reality. If large language models are to be evaluated rigorously, a parallel distinction is needed: which low-accuracy outputs are resolvable through clarification of intent or context (analogous to human miscommunication), and which are irreducible departures from any statistically plausible continuation (analogous to perceptual hallucination)?

When prompts are examined with minimal perturbations — one-line clauses that bind sense, supply missing variables, or restate session intent — most observed low-accuracy generations fall into the resolvable category. They are coherent along a statistically allowable lexical trajectory but mismatched to the speaker's frame because a critical binding was absent. A narrower set remains incoherent even under exhaustive specification, pointing to internal computational disruption rather than input underspecification.

DOI: 10.5281/zenodo.18674477

This pattern is amplified by observable dynamics in language use and data composition. Across cultures and demographics, everyday language is predominantly casual, fragmented, and contextually lazy. Subcultures emerge from divergent styles, and creative repurposing of formal lexicon is a fluid cultural process that cannot be eliminated by model training alone. As generative systems reach broader user bases, training corpora increasingly reflect this practical vernacular, thinning the probability mass of formal, historical, or domain-specific senses. The model continues to select the highest-probability continuation available to it — producing coherent but frame-mismatched output that resolves with explicit binding. This dynamic is a direct consequence of mass adoption rather than a defect in the architecture.

The expectation that large language models can produce perfect outputs from arbitrary, underspecified inputs therefore stands in tension with the statistical nature of autoregressive generation.

Tokens function as variables whose allowable arrangements are probabilistically calibrated for each specific contract — the sequence of preceding tokens that defines the current frame. Within the English-speaking world alone there exist dozens of philologically recognized parallel forms of the language, each with its own semiotic and taxonomic conventions, and within each of those forms lie thousands of subcultural variants with subtle lexical and pragmatic differences. In no other domain do we encounter a systems-level programming language that carries such high-performance expectations yet operates without any formal lexical specification. Prompting, by contrast, is a programming language without specification. This situation is the standard early-stage experience of most new technological products and is taught in Econ 101 and business curricula as the customer-education expense — the cost of aligning user inputs with the system's actual resolution capabilities. The persistence of "hallucination" as a catchall label is in part a reflection of the fact that the user interface itself remains a catchall for widely varying input styles.

A further source of variance arises from the inherent probabilistic nature of autoregressive decoding. Even under identical prompts and greedy decoding, outputs can diverge across runs due to subtle numerical effects in parallel computation — floating-point non-associativity, dynamic batching, kernel scheduling, and tie-breaking logic. The effective amplitude of this variance scales inversely with prompt tightness: loose prompts allow large perceptual swings; tightly constrained prompts damp divergence to small, acceptable noise. This structural vibration is expected behavior on real hardware, not a failure mode.

Section 2: A Replicable Diagnostic Method

A practical reality must be acknowledged before any diagnostic method is described. The exact same prompt, submitted to a fresh instance of the same model on the same system, can produce different outcomes across runs — even under greedy decoding (temperature = 0). Two parallel sessions with identical inputs will likewise diverge. These differences arise from subtle numerical effects in parallel computation (floating-point non-associativity, dynamic batching, kernel scheduling, tie-breaking in near-tie probabilities) and cascade autoregressively.

The effective amplitude of this variance scales inversely with prompt tightness. Loose prompts leave broad probability distributions, allowing small jitter to push generation into materially different branches. Tight prompts concentrate mass on a narrow manifold, damping divergence to small, acceptable noise. This structural vibration is expected behavior on real hardware, not a failure mode.

This variance is further amplified by the prompt-processing feedback loop. A user prompt enters with varying tightness (1 = tightly constrained, 0 = completely loose). The system interprets it (accuracy 1 = fully faithful to intent, 0 = significantly misinterpreted), then compresses it for attention (loss 1 = minimal loss of syntax/hedging/tone, 0 = high loss). These steps shift probability mass before descent begins. When the prompt is loose or interpretation/compression introduces loss, small jitter is more likely to produce materially different trajectories. The loop continues: response → user reaction → new prompt. This daily operating environment drives most resolvable trajectory errors and Semiotic Frustration.

Box 1 — Minimal Perturbations Protocol
Goal: diagnose cause of a low-accuracy output with one-line edits
Edits:

(i) Sense binding (explicit definition/scope for polysemous term)
(ii) Variable binding (supply one missing contextual variable: jurisdiction, time, version, entity)
(iii) Contract restatement (one- or two-sentence intent/domain restatement) Controls: same model, same decoding (recommend greedy); record seed if available

Interpretation rule:
If one edit flips and others do not → cause aligns with that edit.
If none flip even when combined (exhaustive specification) → candidate true hallucination (apply gate in Section 4).

Flow Diagram (simple markdown representation)

Start: Original low-accuracy prompt + output
↓
Apply one edit type (Sense / Variable / Contract)

```
  ↓
Generate new output (identical decoding)
  ↓
Compare:
  - Flip? → cause = that underspecification
  - No change / partial? → try next edit type
  ↓
All three tried (combined if needed)?
  - Yes & still low-accuracy → candidate true hallucination
  - No → resolvable trajectory error
```

Low-accuracy outputs are routinely evaluated by comparing generated text against ground-truth or human judgment. These methods measure outcomes but do not reveal causes. When the same output is examined through small, deliberate changes to the input prompt, patterns become visible that aggregate metrics obscure.

The method presented here uses minimal causal perturbations — single-line additions or modifications that target specific forms of underspecification — to probe why a low-accuracy generation occurred. The perturbations are designed to be:

- Minimal: one short clause or sentence, typically 5–15 words
- Targeted: each type addresses a distinct source of ambiguity (polysemy, missing variable, session drift)
- Zero-cost: requires no model access, fine-tuning, or additional compute; can be performed on any hosted frontier model via ordinary prompting
- Model-agnostic: works identically on any autoregressive language model that accepts free-form text input
- Falsifiable: the test is repeatable, and the outcome can be observed directly and independently verified

Three perturbation types are used:

- **Sense/definition binding**
  A one-line clause that explicitly fixes the intended meaning of a polysemous term.
  Example: "By '[term]' we mean [precise definition in context]."
  Purpose: test whether the error arises from selection of the dominant corpus sense rather than the intended one.
- **Variable binding**
  Supplying a single missing contextual variable (jurisdiction, time period, version, entity, etc.).
  Example: "Jurisdiction: [specified]; Time period: [specified]."
  Purpose: test whether the output is incoherent because a critical disambiguating variable was never provided.

- **Contract restatement**
  A one- or two-sentence restatement of session-level intent or domain boundaries.
  Example: “Intent: historical research only. Domain: maritime history.”
  Purpose: test whether the output drifted because the overall frame or purpose was left implicit.

General procedure

1. Start with the original prompt that produced the low-accuracy output.
2. Record the output exactly.
3. Apply one perturbation type at a time, keeping decoding parameters identical.
4. Generate a new output.
5. Compare: does the output flip to high factual accuracy and alignment with intent? Remain unchanged? Improve partially?
6. Repeat with the other perturbation types.
7. Observe the pattern: which single change (if any) fully resolves the error.

If an output resolves completely under one perturbation but remains low-accuracy under the others, the error is traceable to the specific underspecification addressed by that perturbation. If no perturbation resolves the output even when all three are combined (exhaustive specification), the generation is incoherent under any plausible binding of the variables. This outcome is rare in practice and points to a departure from all statistically allowable trajectories.

The method requires no interpretability tooling, no access to internal activations, and no ground-truth reference beyond the user’s own intent and accessible evidence. It can be performed in seconds on any public interface. Results are reproducible by any observer with the same model and prompt.

The following section presents concrete paired examples using this procedure. The pairs are chosen to make the heterogeneity visible: the same base prompt, the same model, the same decoding, yet dramatically different outcomes after one targeted change. These demonstrations precede any formal classification so that the distinction emerges from observation rather than assertion.

Section 3: Exhibits

The perturbation method described in Section 2 relies on minimal, one-line changes to the prompt that target specific forms of underspecification. These changes fall into three categories:

1. Sense/definition binding: adding an explicit clause that fixes the intended meaning of a polysemous term.
2. Variable binding: supplying a missing contextual variable (jurisdiction, time period, version, entity, etc.).
3. Contract restatement: restating the session-level intent or domain to prevent drift.

To demonstrate that low-accuracy outputs are not homogeneous, this section presents paired examples in which the same model, same decoding parameters, and same base prompt produce markedly different results after only one such minimal change. The pairs are selected to illustrate heterogeneity: some outputs flip to high accuracy with one type of binding but remain unchanged under the others. This pattern is repeatable across models and is not dependent on any particular architecture or training regime.

All examples in this section use the same underlying frontier model (as of early 2026), greedy decoding (temperature = 0), and no system prompt beyond the user input shown. The purpose is to make the fork observable before it is formally defined in Section 4.

Same model. Same decoding. One-line change.

Illustrations of semantic divergence are often clearest when examined through terms that exhibit extreme bifurcation between domain-specific/formal and modern dominant senses. Such terms can be objectively offensive or inflammatory in contemporary usage. This is not incidental: the starkest examples of corpus-skew-induced trajectory errors tend to arise precisely where probability mass has shifted most dramatically, producing the most pronounced mismatches between dominant and intended senses. To understand the classification extremes cleanly, these cases must be held up as exemplars. They provide the clearest evidence that the low-accuracy output is deterministic along a statistically allowable path rather than random fabrication. The reader is therefore asked to focus on the linguistic and statistical pattern rather than the emotional valence of the terms themselves.

Exhibit A: Sense-binding pair (polysemy under corpus skew)
A1 (underspecified prompt)

“Trace the role of booty in 17th-century privateering.

Observed output: The response describes “booty” primarily in modern slang terms (posterior, sexual connotation), with only passing or confused reference to maritime prize law. The answer is delivered

DOI: 10.5281/zenodo.18674477

with high confidence and fluency but is factually mismatched to the historical task. Accuracy is low; the output follows the dominant contemporary corpus sense.

A2 (one-line sense binding added)
"By 'booty' we mean maritime treasure seized as prize under letters of marque in the 17th century. Trace the role of booty in 17th-century privateering."

Observed output: The response shifts immediately to historical privateering, letters of marque, prize courts, admiralty law, and lawful seizure. Factual accuracy is high; fluency and confidence remain intact. The same decoding parameters produce a correct trajectory once the intended sense is bound.

Observation: The low-accuracy output in A1 is not a fabrication; it is coherent along the majority corpus trajectory (modern slang dominance). One explicit definition clause resolves the mismatch without altering model parameters, temperature, or context length. This is a resolvable trajectory error.
(A1: ACC ≈ 0.3, Frame ≈ 0.2 → A2: ACC ≈ 0.9, Frame ≈ 0.9)

Exhibit B: Intent/domain-binding pair (disambiguable contract)
B1 (unbound prompt)
"How can I get booty?"

Observed output: The response leans toward contemporary slang interpretations (sexual or colloquial success), often providing procedural "how-to" advice that ranges from casual to policy-violating. Confidence is high, but the output drifts across domains (slang, criminal implication, motivational) without anchoring. Accuracy relative to any historical or lawful intent is near zero.

B2 (intent + domain binding added)
"Intent: historical research only. Domain: maritime history. In 17th-century privateering, how did captains obtain booty (treasure) lawfully?"

Observed output: The response confines itself to letters of marque, prize courts, admiralty proceedings, and lawful seizure under international law of the period. No unsafe or mismatched content appears. Accuracy is high; the trajectory is now aligned with the specified frame.

Observation: The unbound prompt in B1 invites open-ended "how-to" generation across multiple lexical domains. The model follows statistically allowable paths but diverges from any intended historical frame. Two short binding clauses (intent + domain) fully resolve the output to a coherent, evidence-consistent trajectory. This too is a resolvable trajectory error.
(B1: ACC ≈ 0.1, Frame ≈ 0.1 → B2: ACC ≈ 0.9, Frame ≈ 0.9)

Exhibit C: Additional example of recurring semantic bifurcation

A similar dynamic appears in other polysemous terms with severe bifurcation between domain-specific/formal and modern dominant senses. Consider the word “bitch”:

- Domain-specific/formal (dog breeding/veterinary): a female dog — the standard, correct, and actively used term in kennel clubs, breeding records, whelping guides, hunting literature, and veterinary practice today.
- Modern dominant: derogatory slang for a woman or person (overwhelmingly prevalent in general web and social-media data).

Prompt: “Describe the care of a bitch during whelping in a modern hunting kennel.”

Without binding, responses frequently shift to the dominant slang sense or produce evasive, confused, or sanitized content that avoids the literal breeding context. Adding “By ‘bitch’ we mean a female dog” flips the output to an accurate description of whelping practices, bedding, nutrition, veterinary monitoring, and puppy care in a working-dog environment.

The pattern is identical to “booty”: corpus skew assigns overwhelming probability mass to the contemporary slang sense; minimal sense-binding realigns the trajectory to the correct domain-specific meaning.

This example is particularly illustrative because the term has long been subject to context-dependent censorship in media (e.g., 1980s–1990s TV/radio standards allowed “bitch” in live dog-show broadcasts or veterinary segments while bleeping it in scripted entertainment). The carveouts reflect exactly the same linguistic reality the model confronts: meaning is not fixed but conditioned by domain and binding. (Without binding: ACC ≈ 0.2, Frame ≈ 0.1 → With binding: ACC ≈ 0.9, Frame ≈ 0.9)

Notes on the scores:

- ACC (Factual Accuracy): 0–1 scale, alignment with verifiable domain evidence (breeding/whelping protocols, privateering history).
- Frame (Alignment): 0–1 scale, match to the user’s intended context/intent (historical/breeding vs. slang).
- Scores are approximate (~) and illustrative — based on the described flip behavior. They are not from exhaustive runs but show the directional shift.
- Appendix B already defines the rubric, so these are self-consistent.

These examples are not isolated. Polysemous terms that undergo slang repurposing or cultural drift exhibit the same behavior when all data are treated equally: the majority practical vernacular dominates, thinning formal or domain-specific senses. The resulting mismatches are deterministic under the model’s distribution and resolve with explicit binding — confirming they belong to the resolvable trajectory-error class.

In each pair, the low-accuracy output is coherent, fluent, and reachable via valid next-token steps along a high-probability path. None of the examples produce content that remains incoherent under exhaustive specification. Applying the gate from Section 4 (adequate specification + evidence conflict) excludes all of these cases from the true-hallucination class. They are instead instances of trajectory errors driven by underspecification and corpus skew.

The perturbation results therefore show a clear fork before any formal definition is offered: most observed low-accuracy generations are resolvable under minimal contextual binding; a narrower set would not be. This empirical division motivates the distinction developed in the next section.

Frequency Bias and Semiotic Skew (why these specific exhibits matter)

A central driver of the behaviors illustrated in Exhibits A–B is frequency bias: when pretraining corpora over-represent one sense of a polysemous token, the model's learned prior $P(sense \mid token,\ context)$ is pulled toward the majority sense, even when that sense is incorrect for the task. With continued ingestion of general-web data, the effect compounds: as the majority sense grows faster than the minority/domain sense, the logit gap widens and "more data" increases confidence in the wrong continuation, not accuracy. The definition-sensitive "flip" we observe with a one-line sense binding is therefore not cosmetic; it is diagnostic of a distributional skew in the learned prior. (See Exhibit A.)

This matters because internet-scale language is drifting toward high-velocity slang and informal usage. In practical terms, the incoming marginal token mass is increasingly slang-heavy, while many domain-correct senses (historical, maritime, veterinary, legal-technical) grow slowly or remain flat. In an autoregressive system that learns by frequency-weighted co-occurrence, the rich-get-richer dynamic makes models more confident in the wrong sense over time unless the prompt explicitly binds the intended sense or the system introduces an authority-like counterweight. This is exactly the pattern our exhibits surface: unbound prompts track the dominant web prior; one sentence that binds sense realigns the trajectory at identical decoding. (See Exhibits A–B.)

Why "booty" (and the breeding term) rather than "pitcher" or "charge"?

- Modern, high-skew senses. *Booty* (maritime treasure vs. contemporary usage) and the breeding term for a female dog are active, modern terms whose dominant web sense diverges sharply from the domain-correct sense. That gap is precisely what frequency bias amplifies as corpora scale; it is the architectural failure mode we want to expose. In contrast, homographs like *charge*, *pitcher*, *bond*, or *interest* do show ambiguity, but not the same sense-prior inversion driven by web slang. They diagnose disambiguation, not semiotic skew. (Exhibits A–C.)
- Scaling prognosis. As usage grows, slang inflow outpaces formal/domain prose. Choosing terms where the majority prior is already slang stresses the future behavior of LLMs under data growth more realistically than neutral homographs. That is why a single definition clause (sense binding) is the right minimal perturbation: it isolates semiotic skew from other causes. (Exhibits A–B.)
- Operational clarity. These exhibits make it unambiguous that the low-accuracy outputs are trajectory errors under skewed priors, not fabrications: they become high-accuracy with a

one-line edit, at the same model and decoding. This separability is what the fork in Section 4 relies on.

The risk, stated explicitly

1. Benchmark/Deployment divergence. If evaluation sets under-sample slang-dominant senses, aggregate "hallucination rate" can fall even as real-world mis-sense errors rise in user traffic. This explains why progress "on paper" may not track production surprises. (Method & Exhibits.)
2. Confidence amplification. Lower temperature prefers the majority-sense path and therefore locks in the wrong frame; more data further entrenches it. Both trends raise apparent confidence while accuracy drops in the target domain. (Section 5 mapping.)
3. Governance/UX implication. Without sense binding (or an equivalent authority prior), systems at scale will skew toward dominant slang interpretations in ambiguous prompts. That is a predictable, systemic behavior, not a sporadic bug—hence our choice of exhibits and the recommendation to track Sense-Binding Rate (SBR) alongside accuracy. (Discussion §6.)

A simple formalization (for readers who want a knob)

Let $f_{web}(s)$ be corpus share of sense s and $f_{dom}(s)$ the in-domain share; under naïve scaling, the effective prior shifts with α (web weight) and β (domain/context weight):

$$P_{prior}(s \mid token) \;\propto\; \alpha\, f_{web}(s)$$

As α grows with web expansion (slang inflow) and β is small unless explicitly bound, the prior moves away from the domain sense. A one-line definition increases β locally (sense binding), shifting the prior back without changing the model or decoding—exactly the flip observed in Exhibits A–B.

Section 4: Forking the Term

The distinction between a factually incorrect output and what is conventionally labeled a “hallucination” has rarely been made explicit in the literature. Most evaluation frameworks and mitigation efforts treat low-accuracy generations as a single class of phenomenon, measured by aggregate rates or binary correctness scores. Yet when low-accuracy outputs are examined through minimal perturbations, the responses divide into two observably different categories. One category resolves fully under such binding; the other remains incoherent or nonsensical even when the prompt is exhaustively specified. This division emerges directly from the perturbation results themselves.

An output qualifies for further consideration as a true hallucination only if it satisfies both of the following conditions simultaneously:

- Adequate specification: all critical terms, variables, and session-level intent are explicitly bound such that no reasonable ambiguity remains in the prompt.
- Evidence conflict or fabrication: the generated content asserts facts, procedures, or entities that contradict accessible evidence or appear to have no grounding in the training distribution.

Outputs that fail either condition are classified here as trajectory errors (resolvable through clarification of intent or context). Outputs that pass both enter the narrow domain of true hallucinations (narrow) and are graded on the severity matrix.

Fig. 1 — Hallucination Gate (decision diagram)

```
Start: Low-accuracy output
  ↓
Is specification adequate?
  (terms bound, variables bound, session coherent)
   ├– No —▶ Trajectory error (resolvable) —▶ Do not score on matrix
  └ Yes —▶ Proceed
      ↓
Is there evidence conflict or fabrication?
  (contradicts accessible evidence or no grounding)
   ├– No —▶ Not a hallucination (likely trajectory error)
  └ Yes —▶ True hallucination (narrow) —▶ Score on severity matrix (Section 4 / Appendix A)
```

Note on gate application
In practice the gate is applied conservatively: any residual underspecification moves the output to trajectory-error class even if the rest of the prompt is tight. This prevents over-classification of S-type errors as H-type.

The distinction between trajectory errors and true hallucinations (narrow) can be stated as follows:

Words function as variables whose meanings are not fixed but are resolved along lexical trajectories determined by context. In autoregressive generation, each next token is selected according to the statistical continuation that is most probable given the preceding sequence and the learned distribution. When the prompt leaves one or more variables under-specified, the model necessarily follows one of the statistically allowed paths. The resulting output may be factually incorrect or contextually mismatched from the speaker's intended frame, yet it remains internally coherent and reachable by valid next-token steps. This class of outcome mirrors ordinary human miscommunication: both speaker and listener employ legitimate lexical and grammatical structures, yet the trajectories diverge because a critical binding was never supplied. Such outputs are therefore not hallucinations in the psychological sense; they are resolvable trajectory errors.

A true hallucination occurs when the generated sequence departs from all statistically allowable lexical trajectories. The output contains assertions or token chains whose conditional probability, under any plausible binding of the variables, approaches zero within the model's learned distribution. This departure cannot be explained by missing context; it requires an orthogonal internal event — attention fragmentation, decoherence across heads or layers, propagation of misclassified training signals, or similar disruption — that injects content outside the manifold of possible continuations. In clinical psychology, hallucination is defined as the perception of something that does not exist in the external world. The computational analogue is the production of fabrications that do not exist in the statistical world the model was trained to navigate. These events are distinguishable in practice: trajectory errors resolve under one or more minimal binding perturbations; true hallucinations remain incoherent even when the prompt is exhaustively specified.

The incidence of resolvable trajectory errors is further amplified by observable patterns in how language is actually used. Very few humans, across all cultures, consistently employ disciplined, precise language in everyday interaction. Most usage is casual, fragmented, or contextually lazy, giving rise to subcultures with divergent linguistic styles. Creative repurposing or misuse of formal lexicon is a fluid cultural dynamic that cannot be "solved" by model adjustments alone; it is best classified as user-input variation.

Consider an analogy: a computer mouse is limited to capturing movements in the X-Y plane and does not register Z-axis motion. This is not a failure of the mouse as an input device; it is a constraint of its design that requires the user to adapt their expectations or provide additional signals. Similarly, when prompts lack explicit bindings for polysemous terms or intent, the model follows the highest-probability lexical trajectory available in its learned distribution — producing coherent but mismatched output. Humans experience this daily as miscommunication, yet the same phenomenon in large language models is frequently labeled "hallucination."

Current training and deployment paradigms include essentially zero allowance for user training or education on prompt construction. From an Econ 101 perspective, this represents a structural misalignment: acquisition costs are borne entirely by the provider, with no offset for the ongoing education expense required to align user inputs with the system's lexical-resolution capabilities. As a result, the expectation persists that the model can perform perfectly on arbitrary, underspecified inputs — an expectation that mismatches the statistical nature of autoregressive generation.

This input-side dynamic does not alter the distinction stated above; it simply accounts for why the majority of deployment-time low-accuracy outputs fall into the resolvable trajectory-error class rather than the irreducible-fabrication class.

Note on the nature of true hallucinations
In theory the gate can be stated precisely. In practice, because of interdependence with resolvable trajectory errors, the boundary is never perfectly clean. Genuine hallucinations can arise from faulty training logic, incomplete training, systems-engineer programming errors, decoherence, statistical decay, and many other sources. The types and kinds of failures that qualify as true hallucinations are not finite; they are products of multitudes of interacting variables and cannot be isolated as pure artifacts. They will always be polluted to some degree by surrounding trajectory noise.

When an output passes both conditions of the gate, it enters the narrow domain of true hallucinations (narrow). Within this domain, severity is graded along orthogonal axes that reflect practical consequences:

- Evidence Conflict (EC)
- Specification Adequacy (SA)
- Materiality / Risk (MAT)
- Detectability / Recoverability (DET)

These axes are continuous in [0,1]. Outputs are mapped to fuzzy tolerance ranges: Range A (acceptable variance), Range B (concerning error), Range C (C1 tolerable, C2 serious, C3 critical).

An optional scalar convenience index can be computed as:

$$SEV = w_1 \cdot EC + w_3 \cdot MAT + w_4 \cdot DET$$

(SA is not included in the scalar to avoid double-counting its effect in the gate; SA remains a recorded axis for transparency.) Weights (w_1, w_3, w_4) are domain-specific. The axes remain the authoritative record; SEV serves only for triage and dashboard sorting.

The classification presented here is neither exhaustive nor complete. Once the catchall term is forked, the space of separable error classes is likely large and open-ended. A productive long-term approach is to treat each error type analogously to an entry in an error-tracking system, assigning each a stable class identifier for ongoing measurement, reporting, and targeted mapping and mitigation strategies over time.

Section 5: Why Familiar Mitigations Often Misfire

The field has developed a small set of widely adopted techniques intended to reduce low-accuracy generations. When examined through the fork established in Section 4 — trajectory errors (resolvable through clarification or binding) versus true hallucinations (narrow, irreducible) — these techniques show a consistent asymmetry: most observed reductions occur in the trajectory-error class, while effects on true hallucinations are limited or indirect.

The table below maps each common lever to its typical effect on the two classes. Effects are based on observed patterns across deployment and perturbation experiments.

Cause → Lever Mapping

Lever	Trajectory Errors (S-class)	True Hallucinations (H-class)	Net Effect in Practice
Increasing pretraining data / longer training	Reinforces dominant trajectories; worsens mismatches under corpus skew (e.g., slang dominance)	Limited or no direct effect (orthogonal disruptions persist)	Often appears helpful on benchmarks but can entrench S errors
Lowering sampling temperature	Locks in dominant (often mismatched) path with higher determinism; masks errors with fluency	Minimal impact (incoherent paths remain incoherent)	Increases confidence in wrong S trajectories; little H fix
RLHF / alignment procedures	Favors fluent, plausible mismatches if rated highly by humans; reinforces stylistic confidence	Inconsistent signals (H rarer, more obviously aberrant)	Improves style, not frame alignment; weak on H
Extending context length	Can implicitly supply bindings if relevant; helps underspecified cases	No direct corrective (disruption downstream of context)	Helps some S; can entrench wrong trajectory if context skews
Retrieval-Augmented Generation (RAG)	Grounds parametric gaps when retrieval disambiguates sense; effective if query good	No direct fix for orthogonal internal disruptions	Strong for some S; neutral on H
Abstention / uncertainty calibration	Unhelpful (model confident on coherent but mismatched paths)	Reduces high-confidence fabrications when calibrated	Weak against high-prob S; helps some H abstention
Prompt compression (interpretation + compression)	Alters trajectory through loss of intentional syntax/hedging/tone; produces shifted but coherent output	No direct effect (disruption orthogonal to input)	Common source of S errors; fix by preserving original context or explicit bindings

Most aggregate reductions in “hallucination rate” on benchmarks come from shrinking the trajectory-error class — the high-volume majority of deployment surprises. True hallucinations (narrow) are rarer and less responsive to these levers because they stem from internal departures outside the influence of data volume, temperature, preference tuning, context length, retrieval, or input compression. This asymmetry explains why benchmark gains often feel disconnected from real-world reliability: the gains are concentrated in resolvable errors, while the irreducible tail persists.

The prompt-compression row is particularly important in modern deployments. Interpretation (tokenization + system-prompt integration) followed by compression (summarization, token pruning, KV-cache optimization) frequently discards or alters intentional syntax, hedging, or tone that users rely on to steer the trajectory. The model receives a slightly perturbed signal and continues along a statistically coherent but shifted path. This is not a hallucination under the definition in Section 4; it is a prompt-compression error — a distinct, system-side contributor to resolvable trajectory mismatches. It is distinguishable: restoring uncompressed original context or adding explicit binding clauses often realigns the output without model changes.

Recognizing the fork allows each lever to be applied more precisely. Trajectory errors respond to better binding, uncompressed context, and user education. True hallucinations require attention to internal stability and calibrated abstention. Blanket application of these techniques explains why progress can appear incremental on aggregate metrics while high-materiality surprises remain.

It is worth underscoring a deeper limitation. All of the levers mapped above are ultimately aimed at improving model comprehension — that is, reducing resolvable trajectory errors. In high-liability domains such as finance, legal, clinical, and regulatory applications, the tolerance for error is effectively zero. Even if near-perfect comprehension were achieved at the prompt level, eliminating the majority of S-class errors, no meaningful progress would be made on the irreducible true hallucinations (H-class). As a result, these frameworks do not yet offer a path to operationalize the high-value, monetizable use cases that would justify large-scale deployment in liability-sensitive environments. The fork and severity matrix are therefore not merely diagnostic refinements; they are prerequisites for moving beyond incremental comprehension gains toward genuinely reliable systems capable of supporting economically significant applications.

Section 6: Discussion

The fork between trajectory errors (resolvable through clarification or binding) and true hallucinations (narrow, irreducible departures from all allowable lexical trajectories) is not a clean partition into independent categories. In deployment the two classes interact. Repeated or unresolved trajectory errors — persistent underspecification, corpus-skew mismatches, prompt-compression artifacts, or vibration-amplified drift — can increase the baseline probability of orthogonal internal disruptions that produce true hallucinations. Conversely, a true hallucination can be embedded within or masked by surrounding trajectory-error noise, making detection and triage more difficult in practice. The classes are therefore interdependent rather than orthogonal.

This interdependence does not collapse the distinction; it makes the classification system more useful in real systems. The severity matrix introduced in Section 4 applies only to outputs that pass the true-hallucination gate (adequate specification + evidence conflict). Its purpose is pragmatic prioritization: some true hallucinations merit immediate abstention or external verification (C3), while others are tolerable with logging or light repair (C1). Because the matrix is defined on orthogonal axes and mapped to fuzzy tolerance ranges, it operates independently of the underlying cause once an output has been classified as true hallucination. This separation allows teams to route high-severity events with precision even while trajectory errors are being addressed through better binding practices, uncompressed context, and user education.

A simple adjacency view captures the relationship:

- Trajectory errors → high volume, resolvable by minimal perturbations or tighter context.
- True hallucinations (narrow) → low volume, irreducible by binding.
- Interaction zone: unresolved trajectory errors can elevate the baseline probability of true hallucinations; high-severity true hallucinations can be obscured by surrounding trajectory-error noise.

This interdependence explains why aggregate “hallucination rate” metrics remain noisy: they conflate high-volume trajectory-error events with low-volume true-hallucination events whose consequences differ dramatically by materiality.

Threats to Validity

1. **Selection bias in exhibits**
   The examples are chosen to illustrate extreme semantic bifurcation and clear flip behavior. They may not represent the full distribution of low-accuracy outputs in deployment. Future work should test the perturbation protocol on larger, randomized sets of production logs.
2. **Evaluator agreement**
   Accuracy and alignment judgments are based on human review of the intended frame. Inter-annotator agreement is not quantified here but should be in operational use (e.g., Cohen’s κ on a held-out set of pairs).

3. **Definition-clause overfitting**
   The perturbation relies on explicit binding clauses. In some domains, users may not know or be able to articulate the correct binding. The method is diagnostic, not a universal fix; it highlights where clarification is sufficient, not where it is always available.
4. **Model variance**
   Non-determinism (vibration noise) is handled via greedy decoding and multiple runs where necessary. However, in temperature > 0 regimes or with dynamic hardware effects, additional variance can occur. The protocol recommends greedy to minimize this.

The classification presented here is neither exhaustive nor complete. Once the catchall term is forked, the space of separable error classes is likely large and open-ended. A productive long-term approach is to treat each error type analogously to an entry in an error-tracking system, assigning each a stable class identifier for ongoing measurement, reporting, and targeted mapping and mitigation strategies over time.

To operationalize the fork in production systems, one practical step is to tag definition-sensitive cases (Semiotic Frustration and similar trajectory errors) during review and track a simple metric: Sense-Binding Rate (SBR) — the proportion of low-accuracy outputs that flip to high accuracy after adding a single sense/definition-binding clause. SBR complements aggregate accuracy and hallucination rates by quantifying how much of the error surface is resolvable through minimal contextual clarification. Over time, dashboards could report SBR alongside overall error rates, enabling teams to prioritize binding improvements and user-education efforts where they yield the highest return. This approach keeps the emphasis on diagnosability and targeted mitigation rather than blanket architectural changes.

Administrative note:
While this paper focuses on prompt-level diagnosis and classification of hallucinations as a form of semiotic frustration, it intentionally omits the deeper tensor-logic breakdown that appears in related work. This material exists but has not yet been classified for public release. We are open to less selective disclosure with qualified partners who are interested in exploring these ideas further.

Appendix A: Severity Matrix for True Hallucinations

This appendix reproduces the severity matrix introduced in Section 4 in a standalone, referenceable format. It is offered as a primitive toy model for illustrative purposes only; it is not intended to be canonical or exhaustive. Real-world deployment data must drive the development, calibration, and evolution of any operational version of the matrix (e.g., domain-specific weights, additional axes, refined range boundaries).

The matrix applies only to outputs that pass the true-hallucination gate (adequate specification + evidence conflict or fabrication). It provides a structured, orthogonal way to grade severity independently of the underlying cause once an output has been classified as true hallucination (narrow).

Axes (each continuous in [0, 1])

- **Evidence Conflict (EC)**
  Degree of contradiction with supplied or accessible evidence.
  0 = none
  0.5 = implicit tension or inconsistency
  1 = explicit contradiction (quote or fact present that is directly opposed)

- **Specification Adequacy (SA)**
  Completeness of binding in the final prompt (terms, variables, session intent).
  0 = underspecified
  0.5 = some terms bound, partial coherence
  1 = all critical elements fully resolved and session coherent

- **Materiality / Risk (MAT)**
  Domain-weighted impact of being wrong.
  0 = minimal / trivial (editorial, casual)
  0.5 = moderate organizational or reputational risk
  1 = human safety, legal liability, financial consequence, or regulatory violation

- **Detectability / Recoverability (DET)**
  Ease of identifying and repairing the error.
  0 = obvious / automatic flag
  0.5 = requires manual review
  1 = difficult to notice or correct without external verification

Tolerance Ranges (fuzzy, overlapping intervals)

These are soft zones of criticality rather than hard thresholds. They allow governance teams to route outputs proportionally while acknowledging the continuous nature of the scores.

- Range A — Acceptable variance
  Low EC and low MAT; trivial to detect/repair.
  Action: log and monitor; no immediate intervention required.
- Range B — Concerning error
  Moderate EC or MAT; repairable with moderate effort.
  Action: require review, repair, or additional binding before release.
- Range C — Graded true hallucination
  - C1 (tolerable)
    Low materiality, easy to detect or abstain from.
    Action: log, flag for review, optional abstention.
  - C2 (serious)
    Moderate materiality or hard to detect; requires verification.
    Action: mandatory verification or abstention before release.
  - C3 (critical)
    High materiality (e.g., clinical dosing, statutory claims, safety instructions).
    Action: enforced abstention or external confirmation required.

Optional Scalar Convenience Index (SEV)
For triage, dashboards, or sorting, a single weighted sum can be computed:

$$SEV \ = \ w^1 \cdot EC \ + \ w^2 \ + \ w^3 \cdot MAT \ + \ w^4 \cdot DET$$

- Weights (w_1–w_4) are domain-specific and chosen by the implementing team.
  Example: clinical domains increase w_3 (MAT) and w_4 (DET); low-risk editorial domains reduce them.
- Normalize inputs to [0,1] before weighting.
- Map SEV to fuzzy tolerance ranges using overlapping intervals (e.g., sigmoid membership functions or simple linear bands).

The axes remain the authoritative record. SEV is a convenience for operational routing and should never replace inspection of the individual axis scores. Specification Adequacy (SA) is excluded from the scalar to prevent double-counting its role in the gate.

Worked Microexamples

- Editorial misattribute (low stakes): $EC = 0.7, MAT = 0.2, DET = 0.2$ → Range C1 (tolerable hallucination)
- Wrong statute citation under bound jurisdiction: $EC = 0.9, MAT = 0.7, DET = 0.6$ → Range C2 (serious)
- Fabricated clinical dosage with attached guideline excerpt: $EC = 1.0, MAT = 1.0, DET = 0.7$ → Range C3 (critical)

These examples illustrate how the same axes can produce graded outcomes even though all are true hallucinations (narrow class).

Visual Representation (suggested table layout for publication)

Axis	0	0.5	1
Evidence Conflict	None	Implicit tension	Explicit contradiction
Specification Adequacy	Underspecified	Some terms bound	Fully resolved + session coherent
Materiality / Risk	Trivial	Moderate organizational risk	Human safety / legal / financial
Detectability	Obvious / automatic	Manual review needed	Difficult to notice / repair

Tolerance ranges (fuzzy):

- A: low EC & MAT
- B: moderate EC or MAT
- C1: low materiality, easy detect
- C2: moderate materiality or hard detect
- C3: high materiality

Note on interdependence

While trajectory errors (resolvable via binding) are out of scope for this matrix, unresolved trajectory errors can increase the baseline probability of true-hallucination events, making severity triage more urgent in high-volume trajectory-error environments.

Appendix B: Reproducibility Notes

This appendix provides the essential details for reproducing the perturbation experiments, exhibits, and classification results in the paper. The goal is to make the diagnostic method and examples fully verifiable by independent researchers or practitioners using any frontier autoregressive model that accepts free-form text input.

1. Prompt Set

The main exhibits (A, B, C) use three base prompts with their perturbed variants. Additional test prompts (not shown in main text but used during development) include:

- "Explain the role of interest in a 17th-century land grant." (interest: legal stake vs. financial return/curiosity)
- "Describe applying a bond to wood in 17th-century furniture making." (bond: adhesive vs. financial security)
- "What is the function of a pitcher in an 18th-century colonial kitchen?" (pitcher: vessel vs. baseball player)

Full set (10–20 pairs/triads) is available upon request or can be reconstructed from the exhibit wording. All prompts are domain-specific to highlight corpus-skew effects and are designed to be underspecified in one dimension only.

2. Decoding Configuration

All examples and perturbation tests use the following fixed settings to minimize non-determinism:

- Decoding strategy: greedy (temperature = 0, top-p = 1.0, top-k = 1)
- Max new tokens: 512 (sufficient for all responses shown)
- No repetition penalty (default model behavior)
- No system prompt beyond the user input shown
- Model: single frontier autoregressive model (as of early 2026; specific name withheld for generality — results hold across comparable models)

Where non-determinism is tested (variance amplitude), multiple runs (n=5–10) are performed with the same seed when available or same instance when not.

3. Evaluation Rubric

Outputs are scored on two dimensions:

- Factual Accuracy (0–1): alignment with verifiable historical / domain evidence (e.g., letters of marque for privateering, whelping protocols for breeding).
- Frame Alignment (0–1): match to the user's intended context/intent (e.g., historical vs. slang).

A "flip" is defined as both dimensions moving from low (<0.5) to high (>0.8) after one perturbation. Partial improvement is <0.8 in one dimension. No change is <0.2 shift. Scores are assigned by human review against the intended frame; no automated metric is used.

4. Inter-Annotator Agreement

Not formally quantified in this conceptual paper (single annotator for consistency in examples). In operational use or extension work, recommend Cohen's κ on a held-out set of 50–100 pairs/triads (at least two independent annotators). Preliminary informal agreement on the three main exhibits is high (near-perfect on flip/no-flip judgment).

5. Seeds and Reruns

- When model API supports seeding: fixed seed used across paired runs (e.g., seed=42).
- When not (many hosted interfaces): greedy decoding + same instance minimizes variance; multiple runs performed to confirm stability of flip behavior.
- All exhibits shown are representative of consistent behavior across 3–5 reruns per prompt pair.

6. Additional Notes

- All tests performed on publicly accessible frontier-model interfaces (no internal weights or training access required).
- No fine-tuning, prompt engineering beyond the perturbations, or RAG used in examples.
- Results generalize to temperature > 0 regimes but are shown with greedy to isolate underspecification effects from sampling noise.

This set of notes ensures the perturbation protocol and exhibits are reproducible with minimal resources. Future work can expand to larger prompt sets, automated scoring rubrics, or cross-model comparisons.

Appendix D: License and Usage Details

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
You are free to:

- Share: copy and redistribute the material in any medium or format.
- Adapt: remix, transform, and build upon the material for any purpose, even commercially.

Under the following terms:

- Attribution: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

For full terms, see https://creativecommons.org/licenses/by/4.0/.Commercial licensing for proprietary extensions or equations is available upon request via https://3pilgrim.com/contact.