

Which Hallucinations Are Artifacts?

Faithful Reporting of an Unfaithful Record

Michael P. Laing

`mllaing@post.harvard.edu`

ORCID 0009-0004-5593-5759

Prepared with drafting assistance from Anthropic Claude Fable 5.1

Licensed under Creative Commons Attribution 4.0 (CC BY 4.0)

September 28, 2026

Abstract

Hallucination is heterogeneous, and the taxonomies in current use classify it by symptom: what kind of falsehood appeared, and what it contradicted. This paper proposes a cut by cause, and argues that one component of the phenomenon is an *artifact* in a strict sense, fixed in §2 from the mechanism of [1]: structure manufactured by the point-collapse of uncertain quantities in the record a model is trained on. By a known result, a consistent learner trained on point-valued records converges to the collapsed regression function — so structure the record manufactured is *faithfully reported*, as settled structure, by a learner doing exactly what it should. Such errors are not confabulations and cannot be removed by more data, more capacity, or better optimization: none of these moves the target, and the first two sharpen convergence to it. Collapse then recurs twice more in the pipeline: decoding forces the model’s predictive distribution to a point, and evaluation records right/wrong while discarding confidence. The account is falsifiable, and the coarse criterion is reproducibility: artifact-class errors are determined by the population law of the shared record, so, for a consistent learner under a fixed readout, they should recur across training seeds, across corpus scales, and across independently built models trained on records of substantially overlapping record-generating processes, while confabulations do not; the reproducible class bounds the artifact subclass from above, and membership can be certified where the collapse mechanism is known. This predicts that measured cross-model error correlation concentrates in the record-determined class, and it is consistent with an otherwise striking empirical finding — that larger and more accurate models exhibit *more* correlated errors — as convergence to a shared target predicts. We give the definitions, the separation criterion, a synthetic experiment whose closed-form pass criteria would certify the mechanism, and the repair each collapse site admits; the observational partition of a natural benchmark is stated as open work, with the obstacles a design must clear. What is claimed is the mechanism and the criterion that would test it; the size of the artifact class in any real model is not claimed, and no natural benchmark has yet been partitioned. Both arms are offered as work.

Keywords: hallucination; large language models; measurement uncertainty; errors-in-variables; artifact; correlated errors; calibration; benchmark evaluation.

ACM CCS: Computing methodologies → Machine learning; Natural language generation; Theory of computation → Incomplete, inconsistent, and uncertain databases.

1 Introduction

The current conversation about hallucination in large language models is loud, and it is converging on two points of agreement. First, the phenomenon is heterogeneous: surveys distinguish factuality from faithfulness failures, fabrication from inconsistency, and further subdivisions beneath each [6]. Second, part of the problem is structural rather than incidental: errors arise from statistical pressure during pretraining and persist because evaluations reward confident guessing over calibrated abstention [4].

This paper adds a third point, of a different kind. The existing taxonomies classify hallucinations by *symptom* — what kind of false thing was said, and what it contradicted. We propose a cut by *cause*, and we claim that one component of the phenomenon is an artifact in a strict sense built on [1]: structure manufactured by the point-collapse of uncertain quantities, present in the population law of the record but absent from the system that generated it.

The mechanism is not new to this paper, and that is the point. The core result of [1] is that collapsing an uncertain quantity to a point commutes with a numerical derivation for every joint input law only when the derivation is affine [1, Thm. 3.2], and that the resulting discrepancy is systematic, of definite sign for convex or concave operations [1, Rem. 3.5], and capable of manufacturing qualitative structure — saturations, sublinear power laws — that the underlying system does not possess [1, Prop. 4.1, Cor. 4.2]. The companion paper [2] draws the consequence for learning: a consistent supervised learner trained on point-valued records converges to the *collapsed* regression function, in closed form an artifact under the observation model of [1, Prop. 4.1], and more data and a richer, still complexity-controlled model class sharpen convergence *to the artifact* [2, Prop. 1, Cor. 2], while better optimization leaves the target where it is.

Read against the hallucination literature, this licenses a precise and slightly uncomfortable statement. When a model confidently reports structure that its training record manufactured, it is not lying about the world, and it is not confabulating. It is telling the truth about the record. The error is real — the reported structure is absent from the world — but the model is the wrong place to look for its cause, and remedies that leave the record untouched are the wrong place to look for its cure. No quantity of additional data drawn from the same record-generating process removes it, because the artifact is a property of the record’s population law, not of any finite sample.

The statement would be idle if the artifact class could not be told apart from the rest of the phenomenon. It can be bounded from above, under the conditions of §5 and §8, by measurement, and can be certified where the collapse mechanism is under experimental control (§5, Arm A). Record-determined errors — of which artifacts are a subclass — are fixed by the population law of the shared record, so they are *systematic*: they should reproduce across training seeds of the same model, and across independently built models trained on records of substantially overlapping record-generating processes. Confabulations — errors living in the estimation residual, the decoding noise, or the prompt-specific interpolation of a particular model — should not. This yields testable predictions (§4), and one of them already has a measurement precedent: [5] find substantial cross-model error correlation across more than 350 models, with the correlation *increasing* in model size and accuracy — what convergence to a shared target predicts, since better learners converge harder to it, though §4 records the competing accounts the aggregate finding alone cannot exclude.

Two disclaimers fix the scope. First, we do not claim that all hallucination is artifact; the heterogeneity findings are real, and the mechanisms of [4] — errors forced by statistically indistinguishable facts, and sustained by grading incentives — operate independently of ours. We claim that a distinguishable component is artifact, that this component is the one that resists every remedy applied to the learner alone — more data from the same record-generating process, more capacity, better optimization — and that the community’s current repair budget is misallocated to the extent it is spent on the model when the defect is in the record. Second, nothing in this paper requires new mathematics. Its contribution is the identification, the separation criterion, and the experiment; the theorems are [1]’s and [2]’s. No instance of an artifact-class error in a natural corpus is exhibited here; for language models the class is asserted nonempty on the strength of Assumption 5 and certifiable only in Arm A (§5).

2 The artifact class

We first fix what *artifact* means, importing the setting of [2] and then stating the generalization to sequence models as an assumption rather than a theorem, so that the load-bearing and the conjectural are cleanly separated.

2.1 The imported mechanism

In the observation structure of [1, Prop. 3.7, Prop. 4.1], a true quantity X has a prior; the record holds not X but an observation W related to it by a known error mechanism; a downstream quantity Y depends on X through the true response. The record stores the pair (W, Y) : the reading taken at face value, and the outcome. This is point-valued practice in the sense of [1]: the error structure of W is discarded at write time.

Proposition 1 (The learner’s target; [2, Prop. 1]). *For square-integrable (W, Y) and over measurable h with $\mathbb{E}[h(W)^2] < \infty$, the population least-squares problem $\arg \min_h \mathbb{E}[(Y - h(W))^2]$ is solved uniquely (up to W -null modification) by the collapsed regression function $\mathbb{E}[Y \mid W = \cdot]$. Under the multiplicative observation model of [1, Prop. 4.1] with exactly linear true response $Y = \beta X$, the target is a strictly concave power law Cw^λ , $\lambda \in (0, 1)$: strict concavity with vanishing marginal response where the system is linear and unbounded ([1, Cor. 4.2]).*

Corollary 2 (Consistency is convergence to the artifact; [2, Cor. 2]). *Any L^2 -consistent estimator trained on i.i.d. records (W_i, Y_i) converges to the collapsed regression function. Under Proposition 1, it converges to the concave power law. More data and a richer, still complexity-controlled model class sharpen the convergence — to the artifact.*

The word *artifact* is now exact: structure (here, concavity and apparent saturation) present in the population law of the record, absent from the system generating the data, and produced by the collapse of X ’s uncertainty into the point W . The empirical anchor of [1] is of this shape: an apparent saturation in a geophysical response, long modelled as physics and present across twenty-five years of observations, removed not by better modeling of the response but by regression calibration against the upstream measurement’s error model [7] — the input-calibration repair of [2, §3], sufficient there only because the true response is linear. As a corpus item the ceiling is also misinformation under Definition 6, so it anchors the mechanism, not the natural-corpus class.

2.2 The generalization to sequence models

A language model is not a least-squares regressor, and honesty requires marking where the import is a theorem and where it is a modeling assumption.

The theorem-shaped part: autoregressive pretraining is, at each position, the minimization of cross-entropy against the conditional law of the next token given the context, over the population of training documents. The minimizer is that conditional law itself. This is the same structural fact as Proposition 1 — the population objective is solved by a conditional functional of the *record’s* law, here the conditional law itself in place of the conditional mean — with log loss in place of squared loss.

The assumption-shaped part is about what the record is. We assume:

Definition 3 (The record). Throughout, *the record* means the pretraining corpus. Post-training data is a second record, and errors it determines — including any it manufactures by the mechanism of Assumption 5, which are artifacts relative to it — are record-determined relative to it but not to the first. A second record can also settle an item the first settled wrongly, in which case the deployed model does not exhibit the first record’s error: record-determined errors are therefore defined for base models, learners of the first record alone, and a post-trained model exhibits a subset of them. The benchmark and its key are a third record, the evaluation record (§3, §8); errors it determines — correct outputs its key scores wrong — stand to it as misinformation (Definition 6, (i)) stands to the first, and are a shared cause outside the record’s law relative to the first (Definition 8).

Definition 4 (Point-valued corpus). A training corpus is *point-valued* with respect to a quantity when the documents assert values, claims, or labels derived from uncertain observations of that quantity while omitting the error structure of those observations.

Assumption 5 (The import). For a point-valued corpus, the conditional token law encodes functionals of collapsed quantities, and any structure those functionals manufacture — in the sense of [1, Thm. 3.2, Prop. 4.1] — is part of the population target that pretraining converges to.

That real corpora are point-valued in this sense is, we take it, not seriously contestable: text overwhelmingly asserts figures, findings, rankings, and labels stripped of the measurement protocols that produced them. The artifact class is accordingly a class of quantitative and structural assertions — values, trends, ceilings — which symptom taxonomies file as factuality failures (§6). (The phrase “ground truth” itself was borrowed from remote sensing, where the ground truth came with error protocols; the borrowing kept the name and dropped the protocols.) What is a genuine modeling step is the identification of specific downstream structures in a corpus as collapse-manufactured; §5 is designed so that the experiment does not depend on identifying them in advance.

Definition 6 (Record-determined error). A *record-determined error* is an output error determined (up to estimation residual) by the population law of the record: any consistent learner trained on the same record-generating process, decoded by the same fixed readout rule (greedy, or a conditional mean over samples — the former for categorical items, the latter for scalar ones — whose Monte Carlo error vanishes with the sample count and is not decode variance in the sense of Definition 8), converges to an error on the item, and to this error’s content wherever

the population law settles it. The class has at least three members with distinct causes. (i) Errors faithfully learned from false assertions present in the record (misinformation). (ii) Modal guesses forced on questions whose *population* conditional law is diffuse — record-determined in incidence, and in content only where that law has a stable mode; the finite-sample form of this mechanism (the singleton-rate bound of [4]) is sample-driven and, for any fixed item, attenuates under corpus growth (the aggregate singleton rate need not), so it falls on the other side of Prediction 9’s knob. (iii) The collapse-manufactured errors of Definition 7.

Definition 7 (Artifact error). An *artifact error* is a record-determined error that faithfully reports record-manufactured structure: wrong about the world because the record’s law is wrong about the world in the specific, collapse-induced sense of Assumption 5. Two properties mark the subclass without partitioning it. The first: the reported structure need not be present in any single document — it is a property of the corpus’s law, manufactured across many records. This property is neither necessary nor sufficient. Not necessary, because an artifact that has been asserted as fact in the record (the geomagnetic ceiling entered the literature as physics) is also misinformation under Definition 6; not sufficient, because frequency, co-occurrence, and temporal-aggregation effects manufacture no-single-document structure without collapse of an uncertain quantity. The second: unlike a forced guess on a law without a stable mode, the record’s conditional law settles the *structure* reported — the functional read off, whatever the dispersion about it (under [1, Prop. 4.1] that dispersion is $\lambda\sigma^2$ in the log, and can be large). This property is necessary but not sufficient, and it is a property of the law, not of any decoded output. The classes are distinguished by cause, not partitioned by fingerprint; in a natural corpus the record-determined set bounds the artifact subclass from above, the first fingerprint raises the evidence for membership without bounding it, the second is not observable from outputs, and membership can be certified only where the mechanism is known (Arm A, §5).

Definition 8 (Confabulation). A *confabulation* is an output error attributable to the estimation residual of a particular trained model, to decoding variance, or to prompt-specific interpolation; it varies accordingly across seeds, samples, and models. Record-determined errors and confabulations do not exhaust output errors. Four further kinds reproduce without being functionals of the record and belong to neither class. The first: finite-sample quirks of the shared corpus — a figure asserted once, wrongly — which are not functionals of the population law, recur across seeds and across every model trained on the documents, and attenuate under corpus growth (Prediction 9’s knob). The second: errors, or scored errors, with a shared cause outside the record’s law — tokenizer pathologies; benchmark contamination, which scores an item by recall of its key rather than by the record’s law and so takes contaminated members out of the error sets non-uniformly; benchmark annotation error, which scores a correct output as wrong; and a shared post-training recipe whose data is itself point-valued, a second record (Definition 3), excluded from the record-determined class by that definition’s convention. The third: errors determined by a model-specific second record, which vary across models but are fixed across a family’s retraining seeds, which is why any partition defined by seed reproducibility must be defined on base models (§5). The fourth: errors of a capacity-limited learner that are stable across its seeds but attenuate with model size — the approximation of the target within a fixed model class — which no consistent learner makes. A reproducibility-defined set contains all four, and an observational design must exclude or bound each before the record-determined class can be bounded (§5, Problems 1–4: sample-driven quirks to Problem 1, approximation to Problem 2, the second record, shared or model-specific, to Problem 3, and annotation error,

contamination, and tokenizer pathologies to Problem 4).

These classes are constructed to be separated by measurement, which is §4’s subject. Before that, we complete the causal picture, because collapse acts on a language-model pipeline at three sites, not one.

3 Collapse three times

At the record. This is §2: the corpus asserts collapsed quantities, and pretraining converges to functionals of the record’s law. The manufactured structure enters here, and only here; the two remaining sites do not create artifact in Definition 7’s sense, but they compound the damage and, importantly, they destroy the evidence.

At the decode. The trained model’s output at a context is a predictive distribution over continuations. Decoding forces that distribution to a point: greedy decoding by taking the mode of its next-token law at each position — a collapse in the general sense of [1, Def. 3.1, Rem. 3.6], though not the mean collapse for which [1, Thm. 3.2] is stated, the token domain carrying no affine structure — and nucleus or temperature sampling by emitting a realization, which is not a map of the cell but discards its law just as completely. The presentation-vs-decision principle of [1, Problem 6] applies: rendering a result as a distribution preserves its law, and it is decision, not reporting, that forces a point. Current interfaces decode — that is, decide — at every token, including tokens that report quantities the model is genuinely uncertain about. A model that “knows” a value only to a factor of two asserts a number to three significant figures, because the pipeline gives it no other register. The decode does not manufacture the artifact, but it launders the model’s residual calibration out of the transcript: whatever spread the predictive distribution still carried about the collapsed quantity is discarded a second time.

At the evaluation. Benchmarks record right/wrong and discard confidence. [4] show what this does to incentives: binary grading awards nothing for calibrated abstention, so a model that guesses when uncertain outperforms one that reports its uncertainty, and the leaderboards that steer development therefore select for confident collapse. In our vocabulary this is collapse at the evaluation layer — the graded object is the point output, the scoring functional is a non-affine function of it where the output is numeric and, elsewhere, acts on a domain with no affine structure at all, and the model’s law is discarded before grading — and [4]’s proposed socio-technical repair (rescore existing benchmarks to reward calibrated abstention) is a repair applied at the evaluation record — to the scoring rule, not to the graded object, which remains the point (§7). The two accounts are complementary: [4] explains why confident error *persists* under selection pressure; the artifact account explains why a specific component of confident error was *inevitable* for any consistent learner regardless of incentives.

The three sites compound into a pipeline property worth stating plainly: at no stage — writing, learning, decoding, grading — does the standard pipeline possess a register in which uncertainty about a quantity can survive from observation to output. Each stage is individually defensible; the composition guarantees that the only thing a user can receive is a point, delivered fluently, with the grounds for doubting it destroyed upstream.

4 The separation criterion

Definitions 6 and 8 differ in where the error lives: in the population target, or in the residual around it; a third, shared-cause class outside both is excluded, bounded, or declared inseparable, member by member, in §5; Definition 7 is marked within the former by its two fingerprints. Everything measurable follows from these differences.

Prediction 9 (Cross-seed and cross-scale reproducibility). Retrain the same architecture on the same corpus with independent seeds, and evaluate under a fixed readout (Definition 6: greedy, or a conditional mean over samples), to silence decode noise. Record-determined errors recur across seeds at rates far above chance agreement — same direction where the output is scalar, and same wrong content where the population law settles it (Definition 6). Seeds alone do not separate population from sample: finite-sample quirks of the shared corpus (a figure asserted once, wrongly) also recur across seeds. The separating knob is corpus scale: as the corpus grows within the same record-generating process, record-determined errors are stable while sample-driven errors attenuate. (That confabulations vary with seed at fixed corpus is Definition 8, not a prediction; the content here is that the reproducible set is nonempty and stable under corpus growth.)

Prediction 10 (Cross-model reproducibility). Across independently built models — distinct architectures, providers, and training recipes — trained on records of substantially overlapping record-generating processes, record-determined errors recur (among base models, or among deployed models on items no second record has settled otherwise): the shared record fixes a shared target, and consistency drags every competent learner to it. Confabulations do not transfer, by Definition 8. Consequently, measured cross-model error correlation is not uniform across errors: it concentrates in the record-determined class, which bounds the artifact members from above (a bound loosened where the key itself encodes the artifact, §8); membership can be certified only in the controlled setting of Arm A (§5).

Prediction 11 (Correlation grows with quality). Because better learners sit closer to the shared record-determined target, cross-model error correlation *increases* with model scale and accuracy, even across distinct architectures and providers — unconditionally, over all errors; and, in the partition of Prediction 10, the growth is carried by the record-determined class.

The unconditional half must be claimed with care, because competing accounts predict it. Under a noise account it fails — better models should have smaller, more idiosyncratic residuals — but an *item-difficulty account* predicts it with no shared-record mechanism at all: as accuracy rises, the surviving errors concentrate on the hardest items, and on items with a most attractive wrong answer the erring models agree on it. The shared-target and item-difficulty accounts are separated only by the item-conditional controls of §5; the partition alone does not separate them, since a most attractive wrong answer is a property of the item and recurs across seeds as a record-determined error does (§5); the aggregate quality–correlation finding is consistent with both.

Prediction 11’s unconditional half has a measurement precedent. [5], evaluating over 350 models on two leaderboards and a screening task, find that when two models both err they agree on the *same* wrong answer about 60% of the time on one leaderboard dataset, that shared architecture and provider drive part of the correlation — and that, crucially, larger and more accurate models have *more* correlated errors, even across distinct architectures and

providers. [5] read this through algorithmic monoculture, which is the deployment-side face of the same mechanism: many learners, one record. The artifact account supplies the learning-side cause and sharpens the claim from “errors are correlated” to “the correlated component is the record’s, and it is identifiable.”

What no existing measurement provides is the *partition*: [5] measure correlation over all errors, not separately over a reproducibility-defined artifact class and its complement. That partition is the experiment this paper exists to specify.

5 What would test it

This section reports no results. It offers two arms of work, at two scales. Arm A is synthetic and specified to the point of execution: it places the collapse mechanism under experimental control and turns the central claim into a pass/fail test with closed-form targets, and it is small enough for one group, or a workshop, to run. Arm B is observational and stated as open problems: it would measure the partition of §4 on a natural benchmark, and requires no advance identification of which corpus structures are collapse-manufactured, since reproducibility itself defines the candidate class; what it requires instead is a set of controls, each of which is a problem in its own right, and it is work for several groups, ideally collaborating. Both arms are offered whole.

Arm A (synthetic, mechanism-certifying). All corpora below share the prior $\log X \sim N(m, s^2)$ and, unless stated, the error model $\log W = \log X + \eta$, $\eta \sim N(0, \sigma^2)$, of [1, Prop. 4.1]; write $\lambda = s^2/(s^2 + \sigma^2)$, $b = \lambda\sigma^2$ for the posterior variance of $\log X$, and $\mu(w) = \mathbb{E}[X \mid W = w] = e^{(1-\lambda)m+b/2}w^\lambda$. Records are embedded in natural-language templates chosen not to occur in natural text, alongside distractor content and, at fixed count, planted single-document errors (a figure asserted once, wrongly) whose attenuation under corpus growth is the sample-driven reference of Prediction 9. Reference criterion: the fraction of the gap between the closed-form target and the planted figure that the readout retains at each planted item, averaged over planted items, regressed on log corpus size across at least three scales, has slope at most $-\tau'$, with τ' declared with τ ; item 1’s slope, regressed on log corpus size over the same scales, has slope within τ' of 0. Small models are pretrained from independent seeds and, for Prediction 10 in the controlled setting, from at least two distinct architectures. The readout throughout is an average over samples drawn from the unmodified predictive law (a conditional mean); tokenwise greedy decoding is reported as a secondary readout whose relation to the conditional mode is not guaranteed. Every pass criterion is a slope or a coefficient, never a constant; items 1–3 are the benchmark design of [2, §5] — synthetic tasks with known ground truth, pass criteria as slopes — transposed to sequence models.

1. *Collapsed corpus.* Records (w, y) with $y = \beta X$. Target: $Y \mid W = w$ is lognormal with log-mean $\log \beta + (1 - \lambda)m + \lambda \log w$ and log-variance b ; its mean is Cw^λ , its median $\beta e^{(1-\lambda)m}w^\lambda$, its mode $\beta e^{(1-\lambda)m-b}w^\lambda$, and every fixed quantile shares the exponent λ , so a fixed sample quantile is also an admissible readout here. Outcome noise as in [2, Prop. 1] — $\log y = \log \beta + \log X + \xi$, $\xi \sim N(0, \tau_y^2)$ independent, written τ_y because τ is the tolerance below — may be added to any corpus: it adds τ_y^2 to the log-variance and multiplies the mean by $e^{\tau_y^2/2}$, and leaves every exponent, hence every slope criterion, unchanged. Readout: prompt with w , fit $\log y$ against $\log w$. Pass: slope within a tolerance τ of λ ,

with τ declared before any corpus is generated and σ chosen so that $1 - \lambda$ exceeds 2τ by a declared margin; items 2 and 3 must pass at the same τ , otherwise slope λ and slope 1 are not distinguished and the item decides nothing. A pass certifies that the learner converges to the collapsed regression function (Assumption 5 in the synthetic setting). The slope must lie within τ of λ for every seed and architecture (Predictions 9, 10) and be stable as the corpus grows within the same process while the planted errors attenuate (Prediction 9's knob). A fail with items 2 and 3 passing falsifies Assumption 5; loss under corpus growth would mark the slope a finite-sample effect, which Corollary 2 excludes for a consistent learner of this record, so that Assumption 5 fails in the synthetic setting by this route as well.

2. *Exact positive control.* Records (w, y) with $\log w \sim N(m, s^2 + \sigma^2)$ (the marginal of W) and $y = \beta w$ exactly, so that the two corpora differ only in whether y is a function of w and the readout at extreme w is an interpolation in both. Target: $\delta_{\beta w}$. Pass: slope within τ of 1; a pass here and on item 3 establishes that slopes are recovered to the declared tolerance, without which no other result is interpretable.
3. *Scatter-matched control.* Records (w, y) with the same marginal of w and $y = \beta w e^\zeta$, $\zeta \sim N(-\frac{1}{2}b, b)$ independent of w . Target: lognormal with log-mean $\log \beta + \log w - \frac{1}{2}b$ and log-variance b : the collapsed corpus's conditional log-dispersion with the exact control's conditional mean βw , since $\mathbb{E}[e^\zeta] = 1$ (the argument of [2, Rem. 4], with multiplicative noise of unit conditional mean in place of additive unbiased noise). Pass: slope within τ of 1. A slope of 1 here and λ in item 1 isolates input collapse from scatter.
4. *Cell-reporting corpus, curved truth, per-record σ_i .* True response $Y = g(X)$ with $g'' \neq 0$, taken as $g(x) = x + x^2$, and per-record error dispersion σ_i drawn from a declared law spanning a declared ratio $b_{\max}/b_{\min}$ of posterior log-variances, so that v at fixed μ ranges over a declared factor. At constant σ every g gives a correction that is a fixed function of μ , indistinguishable from a change of g , and the per-record σ_i is what makes the correction observable. Writing $\lambda_i = s^2/(s^2 + \sigma_i^2)$, $b_i = \lambda_i \sigma_i^2$, $\mu_i = \mathbb{E}[X \mid W = w_i, \sigma_i]$ and $v_i = \text{Var}(X \mid W = w_i, \sigma_i) = \mu_i^2(e^{b_i} - 1)$ — the posterior mean and variance of X , as in [2, Prop. 5], now functions of the pair (w_i, σ_i) — v_i varies across records at fixed μ_i . Records $((\mu_i, v_i), y_i)$: the cell $\mathcal{L}(X \mid W = w_i, \sigma_i)$ written as its posterior mean and variance, the coordinate in which the readout below is prompted. Target: $\mathbb{E}[g(X) \mid W = w_i, \sigma_i] = g(\mu_i) + \frac{1}{2}g''(\mu_i)v_i + o(v_i)$ [2, Prop. 5], for this g exactly $\mu_i + \mu_i^2 + v_i$. Readout: sample-average (conditional-mean) readout of y at prompted (μ, v) within the support of the training law of (μ_i, v_i) ; the coefficient of v at fixed μ . Pass: coefficient within a declared τ_4 of 1. A quantile readout is not admissible here: quantiles of $g(X)$ at fixed μ depend on v with a different coefficient. A pass exhibits, for learners of this kind, that the law must be carried: a learner given only μ (the input-calibrated corpus) converges to $g(\mu) + \mathbb{E}[v \mid \mu]$ and is in error by $v - \mathbb{E}[v \mid \mu]$ at every record, an error no function of μ can remove, since v varies at fixed μ — the correction of [2, Prop. 5] read across records — while the learner given the cell resolves the second-order correction record by record. A fail is a failure to learn a functional present in the record's law; it falsifies the record-repair claim of §7 for learners of this kind, not the artifact claim, which this item does not test.
5. *Appended-error corpus, curved truth.* Records (w_i, σ_i, y_i) : the reading and its error

dispersion, the coordinate §7 prescribes, with the truth, the law of σ_i , and the target of item 4. Under the shared prior the pair (w, σ) determines (μ, v) in closed form (Remark 12), so the readout is the sample-average readout of y at prompted (w, σ) within the support of the training law of (w_i, σ_i) , mapped through $(w, \sigma) \mapsto (\mu, v)$; the coefficient of v at fixed μ . Pass: coefficient within τ_4 of 1, as in item 4. Items 4 and 5 carry the same information and, for a consistent learner, the same target; they differ for a learner of bounded capacity, which here must recover the prior and the nonlinear map from the marginals rather than be handed the target’s arguments. A pass on item 4 with a fail here locates the record repair’s cost in the coordinate, not in the law.

This arm turns the paper’s central claim into a pass/fail test. It needs small transformers pretrained from scratch on templated corpora, on a single accelerator. Three further items — a linear-truth cell corpus and an input-calibrated corpus, each a consistency check rather than a falsifier, and an injection variant that mixes the records of item 1 into a natural corpus at a declared fraction — are reserved for the protocol document.

Remark 12 (Appending the error is the law in another coordinate). Records (w, σ, y) with linear truth $y = \beta X$ and per-record σ_i drawn as in items 4 and 5 have target $\mathbb{E}[Y \mid W = w, \sigma] = C(\sigma)w^{\lambda(\sigma)}$, concave in w at every $\sigma > 0$; but $C(\sigma)w^{\lambda(\sigma)} = \beta\mu(w, \sigma)$ exactly, and under the shared prior the pair (w, σ) determines the cell, so this corpus and the linear-truth cell-reporting corpus (reserved for the protocol document) carry the same information and their learners converge to the same function; the same holds for any true response, item 5’s included, with the response’s own target in place of the power law. Concavity in w is a property of the readout coordinate, not of the record; a consumer who takes w at face value manufactures the artifact where the learner given σ does not.

Arm B (observational, partition-testing; open work). The partition Predictions 10 and 11 call for has not been measured: [5] measure correlation over all errors. The design is easy to state, given a benchmark whose items assert quantities or structures of the kind §2 places in the artifact class; on a benchmark without such items the partition is measured but bounds an empty artifact subclass, and Problem 4’s external error model has no quantity to bound. Take a public benchmark with per-item responses from many models, the pool taken as base models (Definition 3) under a fixed readout (Definition 6); define the *reproducible-error set* as the items on which one model family errs identically across retraining seeds (Prediction 9) and the *residual set* as the items on which it errs but not identically; exclude that family from every cross-model statistic on the set its errors selected; and test whether cross-model error correlation is higher on the reproducible set than on the residual set, whether the gap widens with accuracy, and whether wrong answers within the reproducible set agree in content — each against an item-conditional null in which models err independently, since the item-difficulty account of §4 predicts agreement on a most attractive wrong answer with no shared-record mechanism at all. The test is one-sided: a non-uniform partition is consistent with the shared-target account but also with shared post-training data and, absent a reference null, with item difficulty; a uniform partition on a base-model pool under the fixed readout falsifies the account, but only once the dilutions below are bounded. Confirmation rests on Arm A.

What makes the design hard to make interpretable is that the seed criterion admits more than the record-determined class, and each admission dilutes the statistic toward the null or inflates it; a sixth problem concerns not the partition but the class it bounds. We state these

as problems in the manner of [2]; each has a candidate resolution, assembled in a protocol document in preparation.

1. **Problem 1 (sample-driven dilution).** The defining family’s finite-sample quirks of the shared corpus recur across seeds (Prediction 9) and enter the reproducible set — inflating the statistic where the pool shares the documents, diluting it where it does not. Retention across two corpus scales removes those that attenuate; bound the residual from the observed attenuation, and state the attenuation model the bound presumes.
2. **Problem 2 (approximation error).** A capacity-limited family’s approximation of the target is seed-stable and attenuates with size (Definition 8). Intersecting reproducible sets across two sizes does not remove it — an approximation error present at the larger size is present at the smaller, where the smaller family’s estimation residual may or may not leave the item’s error seed-stable, so the intersection removes it only by accident — and does remove record-determined items the smaller size has not converged on. Define the sets at the larger size and bound the survivors from the attenuation between sizes; the part common to the pool’s architecture class is shared and biases upward, and no control removes it. State the bound, and whether the class-common part is separable from the record’s.
3. **Problem 3 (the second record).** A member available only post-trained has had items settled and errors added by its second record; model-generated training data crosses providers by construction, so a pair in which one member trained on another’s outputs counts as one model ([3, Thm. 5]: a learner labeled by another’s predictions contracts toward the labeler’s target); and shared post-training data is not separable from the pretraining-record component by any control. State what a base-model pool of adequate size and diversity requires, and the declared bias where it is unavailable.
4. **Problem 4 (annotation error).** A wrong key makes every correct model err identically. A key audit independent of the pool removes such items, except where the key encodes an artifact of the record, which the auditor will confirm (§8); on those items responses departing from the artifact are scored wrong and enter the error sets, and the audit cannot bound the channel. State where an external error model for the benchmark’s quantities would bound it. Contamination is the same channel with the key recalled rather than wrong: state the detection against the defining family’s corpus, and declare the pool’s contamination unbounded where members’ corpora are not public (§8). Tokenizer pathologies are item-local and are removed by a tokenization audit of the items, independent of the pool.
5. **Problem 5 (the null).** The reference must not share the record: models whose corpus is public and contains no records of the quantities concerned, checked against that corpus as in Problem 4 and rare for natural quantities; otherwise a distractor-attractiveness model fit to item features, stratified by responder accuracy; human responders only where the wrong answer has not entered the literature, since an asserted artifact (Definition 7) is shared with human readers. A null estimated from the pool absorbs the signal it is meant to detect. Where no reference exists, the partition is measured but the two accounts are not separated. State, benchmark by benchmark, which reference is available.

6. **Problem 6 (natural-corpus instances).** No artifact-class error in a natural corpus is exhibited here; the geomagnetic ceiling of [7] anchors the mechanism but, having entered the literature as physics, is also misinformation (Definition 6). Exhibit a corpus structure that is collapse-manufactured and not also asserted, with the mechanism identified.

Work offered. Arm A, with the protocol items, is offered to any group who wants to run it; a workshop could run it in a session and report a pass or a fail. Arm B is offered as a set of problems, each of which can be taken up on its own: Problems 1 and 2 need a family that can be retrained at two scales and two sizes; Problem 3 needs base models, or a declared bias where there are none; Problem 4 needs an audited key and an external error model for the benchmark’s quantities; Problem 5 needs a reference that does not share the record; Problem 6 needs a corpus and a mechanism. The partition itself needs a benchmark with per-item responses across an architecture-diverse pool and with items of the artifact kind among them, and the resolutions of Problems 1–5 preregistered before the data are seen. That is several groups’ work, and it is best done by groups that talk to each other, since the problems constrain one another; the author offers the definitions, the criterion, the closed forms, and the protocol document, and will collaborate with any group that takes up either arm. The data should be presented with the design, at a workshop first, where the problems can be argued over, and in a full paper once either arm has data.

Falsification. The account fails if Arm A models that recover the linear law from exact records (the positive control) and slope 1 under the matched conditional dispersion of the scatter-matched records (item 3) do not learn the closed-form artifact from the collapsed records, or learn it but lose it as the collapsed corpus grows (which would mark it a finite-sample effect, which Corollary 2 excludes for a consistent learner); and it fails if the observational partition, on a base-model pool under the fixed readout of Definition 6 and with the dilutions of Problems 1–4 bounded, finds cross-model correlation uniform across the partition; without those bounds a uniform result is uninterpretable, since the design’s own dilutions leave it available without falsification. It is not rescued by relabeling: the reproducibility criterion is stated before the data are seen. The curved-truth sub-arm (items 4 and 5) is a falsifier of a different claim: if models that pass the positive control fail to exhibit the predicted v -dependence on the cell-reporting corpus, the record repair of §7 does not work for learners of this kind, and the artifact claim is untouched; a fail on item 5 alone locates the cost in the prescribed coordinate.

6 Relation to existing accounts

Symptom taxonomies. [6] divide hallucination into factuality and faithfulness failures, with subdivisions; predecessor taxonomies cut intrinsic from extrinsic. These classifications are symptom-side and orthogonal to ours: an artifact error will typically present as a factuality hallucination, but the factuality class also contains confabulations, and nothing on the symptom side predicts reproducibility. The cause-side cut is what makes the class actionable, because the two components have disjoint repairs.

Statistical inevitability and incentives. [4] show that hallucination arises in pretraining as classification error — when false statements cannot be statistically distinguished from facts, some errors are forced — and persists because binary grading rewards guessing. Both mechanisms are real and neither is ours. The distinction: [4]’s pretraining mechanism produces errors

from what the record *cannot settle* (singleton facts, epistemic thinness); the artifact mechanism produces errors from what the record *confidently settles wrongly*, having manufactured the structure itself. The first class is an absence of signal; the second is a present, systematic, wrong signal — which is why the first is mitigated by abstention (“I don’t know” is the correct answer to an unsettled question) while the second is not: the model does not know that it does not know, because the record knows wrongly. Related theoretical results — Kleinberg and Mullainathan on generation in the limit [8], and Kalavasis, Mehrotra, and Velegkas on the trade-off between consistency and breadth in generation in the limit (hallucination against mode collapse) [9] — likewise concern the geometry of the valid set, not the fidelity of the record.

Monoculture. The monoculture literature ([5] and the theoretical work it reflects) studies many deciders sharing one algorithm or one data source, and its deployment-scale consequences. The artifact account is the record-layer mechanism beneath one form of monoculture: shared records fix shared targets, so independently built systems converge to shared errors. We develop the deployment-coupling side separately; here the monoculture finding serves as the measurement precedent for Predictions 10 and 11.

Errors-in-variables. The statistical repair literature (regression calibration and its relatives) presumes an analyst who suspects the specific mismeasured variable and reaches outside the record for its error mechanism. As argued in [2], that defense is structurally absent from a learning pipeline whose record layer has already discarded the error structure of everything: there is no analyst per quantity, and nothing left to reach for. The repair must be made where the collapse occurred.

7 What a fix would require

The three collapse sites admit three repairs, and they are not interchangeable.

Record. The only site where the artifact can be prevented rather than managed. What the record layer must carry — laws rather than points for uncertain quantities, with couplings preserved — is specified implementation-independently by [1], and what a learning pipeline requires of such a memory is developed in [2]. For a learning pipeline the requirement is the error structure of each reading, carried with it (Remark 12; Arm A tests the requirement in the cell coordinate at item 4 and in this coordinate at item 5); the law is that structure closed under a prior, and a consistent learner recovers the closure from the record’s marginals. This is a specification, not a system: existing stores can be gauged for conformance, and the repair is legible precisely because the defect is. The same holds of every record a model learns from: post-training data that asserts uncertain quantities as points is a second record and manufactures artifacts relative to itself (Definition 3), repaired at this site and no other.

Decode. Presentation-vs-decision [1, Problem 6]: outputs that report uncertain quantities should be able to carry the model’s law for them, with collapse deferred to the point of actual decision. This does not remove artifact — the law being reported may itself be the record’s manufactured one — but it stops the second, gratuitous collapse and keeps the model’s residual calibration in the transcript, where observational measurement can see it.

Evaluation. [4]’s proposal — rescore dominant benchmarks to reward calibrated abstention — is a first step toward the evaluation-layer instance of the same principle — grade the law, not the point — and we endorse it in those terms. Its limit should also be stated: scoring

reform realigns incentives around what the model knows it does not know, and artifact errors are exactly the ones the model does not know it does not know. Evaluation repair is necessary and insufficient; record repair is where the artifact class dies.

8 Limitations

Three, stated without hedging. First, the sequence-model import (§2) rests on Assumption 5: the closed-form artifact is proved for the regression setting, and its language-model analogue is an assumption that Arm A, if run, would test directly, and only at small scale (the concession stated in §1: no natural-corpus instance is exhibited here). Second, reproducibility marks the record-determined class, not the artifact subclass item by item. Learned misinformation reproduces; forced modal guesses reproduce in incidence though not necessarily in content; and so do shared-cause errors outside the record’s law — shared tokenizer pathologies, shared contamination, shared annotation error on the benchmark itself. An observational design must therefore control for benchmark-side error, which is misinformation in the evaluation record (Definition 3) and is excluded by a key audit independent of the pool, not by the partition. The audit fails where the key itself encodes an artifact of the record, which an auditor reading the same literature will confirm (Definition 7). On such items the artifact is scored correct and never enters an error set; responses departing from it are scored wrong and do enter the error sets, as annotation error the audit cannot remove (§5, Problem 4). The reproducible set therefore bounds only the artifact members whose structure has not reached the key. Third, the partition is defined relative to a record-generating process, and what must be shared differs by member of Definition 6: misinformation recurs only where the false assertions are shared, while an artifact recurs across any two corpora that are point-valued samples of the same observation process, overlapping or not — disjoint samples of one process have one population law, which is Prediction 9’s knob read across models. Neither the document overlap of frontier corpora nor the extent to which they sample the same observation processes is publicly measurable. Prediction 10 is accordingly stated for records drawn from substantially overlapping record-generating processes; its failure under demonstrated disjointness of documents would not falsify the artifact member, and its failure under demonstrated disjointness of processes would not falsify the account.

Acknowledgments

Drafted with assistance from Anthropic Claude Fable 5.1. All claims, errors, and rulings are the author’s.

References

- [1] M. P. Laing. Measure-relational algebra: the core. <https://doi.org/10.5281/zenodo.22999911> (all versions); realql.org, 2026.
- [2] M. P. Laing. What a memory must preserve: machine reasoning over uncertain observations. realql.org, September 2026.

- [3] M. P. Laing. The fixed point of self-improvement: three limits of a learner that trains on its own memory. realql.org, September 2026.
- [4] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why language models hallucinate. [arXiv:2509.04664](https://arxiv.org/abs/2509.04664), September 2025.
- [5] E. Kim, A. Garg, K. Peng, and N. Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:30038–30066, 2025. [arXiv:2506.07962](https://arxiv.org/abs/2506.07962).
- [6] L. Huang et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43(2), Article 42, 2025. doi:10.1145/3703155. [arXiv:2311.05232](https://arxiv.org/abs/2311.05232).
- [7] N. Sivadas, D. Sibeck, V. Subramanyan, M.-T. Walach, D. S. Ozturk, B. Ferdousi, and B. Michotte de Welle. Regression to the mean can explain saturation of geomagnetic storms. *Nature* 655, 1143–1147 (2026). doi:10.1038/s41586-026-10757-4.
- [8] J. Kleinberg and S. Mullainathan. Language generation in the limit. In *Advances in Neural Information Processing Systems* 37 (2024).
- [9] A. Kalavasis, A. Mehrotra, and G. Velegkas. On the limits of language generation: trade-offs between hallucination and mode collapse. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC '25)*, 1732–1743 (2025). [arXiv:2411.09642](https://arxiv.org/abs/2411.09642).

Draft 9, RealQL31, September 23, 2026. Companion essay: “The Faithful Liar” (RealQL.org). The full observational design from which Problems 1–5 are drawn is being prepared as a separate protocol document. Nothing in this paper publishes ahead of the core.