A Protocol-Fixed Information-Theoretic Reporting Framework for Evidence-Weak Hallucination in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

This paper introduces a protocol-fixed information-theoretic reporting framework for evidence-weak hallucination comparisons in large language models (LLMs). The framework uses sequence-level information energy, variational free energy, KL decomposition, and coarse-graining-induced information loss to define a closed reporting object for controlled evidence-weakness comparisons. Its contribution is the protocol-fixed closure of that object, rather than a new free-energy identity, KL theorem, or scalar hallucination score. Within a declared protocol block, it fixes the admissibility conditions for the comparison coordinates $(\Delta E, T_{\text{eff}}, c_0)$, the `apply/not-apply` gate, the relative identifiability convention, and a mechanically auditable artifact interface. The fully observable synthetic toy implementation remains the principal anchor for showing that the declared comparison object can be instantiated, logged, filtered, aggregated, and assigned an explicit status under controlled evidence weakening. In addition, a black-box Claude 4.5 compression–expansion case study is added as a practical instantiation of the same artifact-level claim-audit convention. That case supports portability of the reporting object to real LLM outputs and records a small factualization drift, but it is not a completed real-LLM validation study, an Arrhenius-rate validation, or a protocol-free law. The result is a bounded reporting framework for auditable, comparable, and explicitly stoppable evidence-weakness comparisons under a declared protocol. It is a reporting framework rather than a deployment-time detector or mitigation method.

1 Introduction

This paper gives a protocol-fixed framework for evidence-weak hallucination under controlled evidence weakening, organized by an information-theoretic comparison backbone. The fully observable toy implementation remains the main procedural anchor: it shows that the declared comparison object can be fixed, logged, filtered, aggregated, and assigned an explicit status before interpretation. To address practical instantiation beyond a synthetic setting, the revised manuscript also adds a black-box Claude 4.5 compression–expansion case study in Section 5. That case is reported as a portability and claim-audit specimen, not as a completed real-LLM validation study or an empirical law for Claude 4.5 behavior. Appendix-level material is retained only as a same-form portability record under the same reporting convention and does not enlarge that main evidential basis.

The practical problem addressed here is not how to detect hallucinations at deployment time, but how to make evidence-weakness comparisons interpretable and auditable. Existing studies often report error rates, uncertainty scores, retrieval conditions, or diagnostic curves under partially different experimental conventions. Such quantities are difficult to compare unless the evidence-degradation rule, external label, logged observables, exclusion rule, and stopping rule are fixed as one object. The present framework formalizes that

The authors used large language model tools for language editing, consistency checking, and formatting assistance. All ideas, claims, mathematical content, experimental design, and final responsibility remain with the authors.

object. Its value is therefore methodological: it specifies when a reported comparison is interpretable, when it is only diagnostic, and when interpretation must stop.

Boundary to related work. The framework is closest in topic to hallucination detection and factuality evaluation, uncertainty-based approaches such as semantic entropy, and retrieval-augmented generation Ji et al. (2023); Maynez et al. (2020); Lin et al. (2022); Min et al. (2023); Manakul et al. (2023); Li et al. (2024); Farquhar et al. (2024); Lewis et al. (2020); Karpukhin et al. (2020); Izacard and Grave (2021); Huang et al. (2025). The distinction is not that this paper proposes a new uncertainty statistic or a stronger factuality label. Existing factuality, hallucination-detection, uncertainty, and retrieval-augmented methods primarily define scores, labels, interventions, or evaluation procedures for model outputs. In contrast, the present framework fixes the entire comparison object: the evidence-degradation coordinate, external answer-segment label, decision-critical diagnostics, invalid-run exclusion rule, and stopping rule are reported together or not interpreted at all. The claimed novelty is this protocol-fixed closure of the reporting object, not a new scalar hallucination score or a deployment-time detector. This distinction is operational rather than terminological: the framework could in principle be used to audit how an existing uncertainty score, factuality label, or retrieval condition is reported under evidence weakening, but it does not replace those methods or claim to improve their predictive accuracy.

Relation to prior theory and disciplinary boundary. The information-theoretic and free-energy components used in this manuscript are standard rather than new: sequence-level information energy is a negative log-likelihood reparameterization, the variational free-energy identity is a KL decomposition, and the coarse-graining statement uses the usual data-processing direction. Classical Arrhenius/Kramers and Mori–Zwanzig terminology is used only to organize local within-protocol comparisons after these quantities have been fixed. The disciplinary boundary is therefore explicit: the paper does not derive a new stochastic dynamics for LLM decoding, but defines a closed reporting object that specifies when such imported coordinates may be reported, when they are only diagnostic, and when interpretation must stop.

Core contribution. The contribution is a closed reporting object for evidence-weakness comparisons. The object closes four components under one declared protocol: the information/free-energy and coarse-graining backbone, the admissibility-and-stopping gate, the relative identifiability convention, and the auditable artifact interface. This closure makes the comparison reproducible at the level of the declared protocol: the evidence-degradation rule, external label, logged diagnostics, invalid-run rule, aggregation rule, and `apply/not-apply` outcome are fixed together. The contribution is therefore methodological: it fixes the conditions under which evidence-weakness comparisons can be reported, audited, compared within a protocol, or explicitly stopped.

Evaluation target and scope. The manuscript should therefore be evaluated as a protocol specification and reporting framework, not as a hallucination detector, benchmark study, or completed real-LLM validation paper. Its scope is limited to five claim-bearing objects: the information/free-energy and coarse-graining backbone of Section 2; the explicit `apply/not-apply` admissibility and stopping gate of Section 3; the relative identifiability convention for within-protocol reporting; the auditable artifact interface checked on the fully observable toy anchor in Section 4; and the black-box claim-audit case study in Section 5. The relevant evaluation questions are whether this declared comparison object is well defined, mechanically auditable where the required artifacts exist, portable to a real-output audit record, and equipped with a stopping rule when its interpretation conditions fail.

Definition of "hallucination" in this manuscript. Throughout, "hallucination" means *evidence-weak* errors under a given input context: the model generates fluent content that is unsupported by, or contradictory to, the available evidence in the realized effective context C_e. In retrieval-augmented settings, retrieval failure is treated only as an upstream cause that changes what evidence enters C_e; here we condition on the realized C_e and vary evidence availability only through the controlled knob L under a fixed protocol (Definition 4.1, Remark 3.7).

Table 1: Minimal notation used throughout the main claim path.

Symbol	Meaning (fixed main-line usage in this manuscript)
$E_\theta(x)$	Sequence-level information energy: $E_\theta(x) = -\log P_\theta(x)$ (Definition 2.1)
$E(x)$	Local continuous surrogate energy on a projected coordinate (Section 3)
ΔE	Surrogate barrier height: $\Delta E = E(x_s) - E(x_m)$ (Section 3)
$\Delta E(L)$	Modeled surrogate barrier coordinate as a function of effective context length L (Section 2.3)
T_{eff}	Analysis-level effective temperature used in free-energy and surrogate dynamics; distinct from τ
β	Inverse effective temperature in the surrogate Langevin/Fokker–Planck setting: $\beta = 1/T_{\text{eff}}$
τ	Implementation softmax temperature used by the fixed decoding protocol
c_0	Prefactor/rate-scale factor in the Arrhenius organizer under fixed protocol
L	Effective context length, $L \in [0, 1]$ (Definition 4.1)
$C,\ C_e,\ \widetilde{C}$	Full context, realized effective context, and generic coarse-grained context
$P_{\text{err}}(L)$	Externally labeled answer-segment error rate under fixed protocol
$g_t,\ \mathcal{H}_t$	Observable diagnostics: logit gap and entropy on decision-critical steps
$\bar{g}(L),\ \bar{\mathcal{H}}(L)$	Decision-critical averages of g_t and $\mathcal{H}_t$ at each L

Guide to the paper. Section 2 fixes the information-energy, free-energy/KL, and coarse-graining backbone. Section 3 fixes the admissibility and stopping conditions, with prefactor bookkeeping limited to the role of c_0 in the comparison shell. Section 4 fixes the auditable toy-anchor protocol. Section 5 adds a black-box Claude 4.5 compression–expansion case study under the same artifact-level claim-audit convention. Appendix A contains only appendix-level non-core material under the same reporting convention.

Notation map. Table 1 is intentionally restricted to the symbols needed on the paper's main claim path.

2 Information and Free Energy

This section fixes the manuscript's information/free-energy backbone: sequence-level information energy $E_\theta(\cdot)$, the variational free-energy functional, the KL decomposition, and the information-loss hierarchy induced by coarse-graining. Within that backbone, T_{eff} is introduced as an analysis parameter, while continuous landscape language is deferred to the local comparison shell of Section 3.

Definition 2.1 (Information energy of an LLM)**.** Let $x = (x_1, \ldots, x_T)$ be a token sequence and let θ denote the model parameters. The *information energy* of x under an autoregressive LLM is defined as

$$E_\theta(x) := -\log P_\theta(x), \tag{2.1}$$

where

$$P_\theta(x) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{<t}). \tag{2.2}$$

Since $E_\theta(x) := -\log P_\theta(x)$, we have $e^{-E_\theta(x)} = P_\theta(x)$. Hence the induced Gibbs-form representation is a reparameterization of the model distribution:

$$p_\theta(x) \equiv P_\theta(x) = \frac{1}{Z(\theta)} e^{-E_\theta(x)}, \qquad Z(\theta) = \sum_x e^{-E_\theta(x)} = \sum_x P_\theta(x) = 1. \tag{2.3}$$

Equation (2.3) is the unit-temperature Gibbs-form reparameterization of the original autoregressive distribution induced by Definition 2.1. The T_{eff}-deformed distribution $p^*(x)$ introduced later in Proposition 2.2 is an analysis distribution, not the original LLM distribution unless $T_{\text{eff}} = 1$ after the chosen normalization. No sampling-temperature or decoding-temperature claim is made from Equation (2.3).

2.1 Variational Free Energy and KL Structure

In this subsection, T_{eff} is an analysis-level effective temperature used only in the analysis functional and kept distinct from the implementation temperature τ unless explicitly specified.

Remark (Analysis-side status of T_{eff}). T_{eff} is introduced here only as an analysis-side coordinate in the variational functional. It is not identified with the implementation temperature τ or any other decoding parameter. Its reportability is fixed later by the Arrhenius-shell reporting rule and the `apply/not-apply` gate.

Proposition 2.2 (Variational free energy and KL decomposition)**.** *Let $q(x)$ be any probability distribution over sequences and define*

$$F[q] := \mathbb{E}_q[E_\theta(x)] - T_{\text{eff}}\, H[q], \qquad H[q] := -\sum_x q(x) \log q(x). \tag{2.4}$$

Let $p^*(x) = Z_{\text{eff}}(\theta)^{-1} \exp\{-E_\theta(x)/T_{\text{eff}}\}$. Then

$$F[q] = -T_{\text{eff}} \log Z_{\text{eff}}(\theta) + T_{\text{eff}}\, \mathrm{KL}(q \parallel p^*). \tag{2.5}$$

In particular, minimizing $F[q]$ over q is equivalent to minimizing $\mathrm{KL}(q \parallel p^*)$.

Integrated backbone fixed in Section 2. The mathematical identities in this section are standard. Their role here is not to supply a new free-energy theorem or a new KL theorem, but to fix the variables and admissible comparison units that are later passed to the protocol-fixed reporting shell. The physics vocabulary therefore has a bookkeeping and organizing role unless the later admissibility gate explicitly permits a stronger interpretation. In particular, sequence-level information energy, the variational free-energy/KL relation, the coarse-graining-induced information-loss ordering, and the analysis-side status of T_{eff} are fixed here so that later `apply/not-apply` decisions have a declared theoretical object to refer to.

2.2 Information Loss under Context Coarse-Graining

We reserve C for the full context, C_e for the realized effective context after truncation, and $\widetilde{C}$ for a generic coarse-grained context. Coarse-graining is treated first as an information-processing map: rigorous information-loss directions constrain what can increase under $C \to \widetilde{C}$ Cover and Thomas (2006). In the standard data-processing sense, for any downstream variable Y under the induced Markov relation $Y \leftrightarrow C \to \widetilde{C}$,

$$I(Y; \widetilde{C}) \le I(Y; C).$$

This information-loss ordering is part of the paper's main backbone. By contrast, any link from information loss under coarse-graining to an effective barrier dependence is treated strictly as a modeling/organizing choice rather than a theorem; no monotone barrier identity is inferred from the data-processing inequality.

2.3 Barrier Model for Context Length

This subsection records $\Delta E(L)$ only as a monotone, saturating modeling input rather than as an identity. Its role is only to supply a local modeling input passed downstream to the Section 3 comparison shell. The two forms below are admissible illustrative parameterizations of a monotone saturating barrier coordinate; they are not required by the framework, not fitted in the main claim, and not used to infer a protocol-free law.

$$\Delta E(L) = \Delta E_{\min} + \alpha \log\bigl(1 + \kappa L\bigr), \tag{2.6}$$

$$\Delta E(L) = \Delta E_{\min} + \Delta E_{\text{span}}\bigl(1 - e^{-\lambda L}\bigr). \tag{2.7}$$

3 Local Comparison Shell with an Arrhenius-Form Organizer

This section adds a local admissibility-and-stopping shell to the information/free-energy backbone fixed in Section 2. The Arrhenius-form expression is used only as a within-protocol organizer, not as a derived dynamical law for token decoding. Together with Sections 2 and 4, it gives the manuscript a three-layer structure: theory, admissibility/stopping, and auditable implementation. Reference points are classical rate theory, rare escape, and Mori–Zwanzig coarse-graining Kramers (1940); Langer (1969); Hänggi et al. (1990); Freidlin and Wentzell (2012); Bovier and den Hollander (2015); Schuss (2010); Mori (1965); Zwanzig (1961).

3.1 Formal commitments of the comparison shell

Section 3 records three formal commitments of the comparison shell rather than theorem-like novelty claims.

1. *Reporting rule.* The Arrhenius organizer is a within-protocol reporting shell, usable only on the admissible no-fail regime.
2. *Compatibility/admissibility rule.* The logged observables are compatibility tests for the declared shell and do not define the latent comparison variables.
3. *Stopping rule.* Once a formal failure condition is triggered, interpretation stops and the correct output is `not-apply`.

Rule 3.1 (Reportability of the protocol-fixed organizer)**.** *Fix one decoding-and-logging protocol $\mathcal{P}$ together with the implementable protocol objects of Section 4; in particular, the model, tokenizer, decoding rule, prompt template, labeler, logging scope, degradation rule, and L-grid are fixed. Write the protocol-level organizer in the Arrhenius shell*

$$P_{\mathrm{err}}(L) \approx c_0 \exp\left(-\frac{\Delta E(L)}{T_{\mathrm{eff}}}\right). \tag{3.1}$$

Equation (3.1) *is not fitted in the manuscript and is not inferred as a protocol-free law; it is used only as an organizing form for the declared within-protocol comparison. With $(\Delta E, T_{\mathrm{eff}}, c_0)$ interpreted only as within-protocol comparison coordinates, this organizer is reportable as a minimal formal shell for within-protocol interpretation only on the admissible no-fail regime*

$\mathcal{P}$ is fixed, *Equation* (3.2) *holds,* *Assumption 3.3 holds,* *and Definition 3.4 is not triggered.*

Outside that regime, the manuscript makes no organizer-level interpretation claim for $(\Delta E, T_{\mathrm{eff}}, c_0)$.

Reporting lock for T_{eff}. To eliminate post-hoc degrees of freedom, T_{eff} is not treated as an intrinsic model temperature. It may be reported only as a within-protocol relative comparison after the L grid, reference level, zero-correction rule if needed, and barrier-unit convention have been fixed before aggregation. If the barrier scale and temperature-like coordinate are both fitted from the same $P_{\mathrm{err}}(L)$ curve, T_{eff} is not reported as independently identified; the manuscript then reports only the observed curves and the declared `apply`/`not-apply` status. Appendix A is curves only and reports no T_{eff} estimate; the black-box case study of Section 5 has no logits and therefore also reports no rate-coordinate estimate.

3.2 Discrete decoding and the projected coordinate

Autoregressive decoding under a fixed protocol induces a stepwise stochastic evolution of an internal state. We introduce a task- and analysis-dependent projection $x = \Pi(s) \in \mathbb{R}^d$ only as a local comparison coordinate, not as a derived state equation. The projected diffusion picture is therefore used solely as a minimal organizational coarse-graining in the Mori–Zwanzig spirit Mori (1965); Zwanzig (1961), only within Assumption 3.3, outside Definition 3.4, and under the fixed-protocol reporting convention of Remark 3.7.

Remark 3.2 (Scope of rate-theory vocabulary)*.* Reaction-rate vocabulary is used in this manuscript only as an organizer for within-protocol comparisons under controlled evidence changes. We use the numbered Arrhenius/Kramers form in Equation (3.1) to structure protocol-controlled comparisons of answer-segment error

rate, with diagnostics computed on decision-critical steps (Section 4), and we keep prefactor bookkeeping explicit via c_0. We do not claim that discrete token decoding obeys a protocol-free Langevin/Fokker–Planck law, and we do not identify $(\Delta E, T_{\text{eff}}, c_0)$ as intrinsic model properties independent of protocol.

Assumption 3.3 (Admissibility checklist for the local-surrogate organizer)**.** The Arrhenius/Kramers statements in this section are used only as a local diffusion surrogate for token-step decoding, and only when the following admissibility conditions are satisfied under a fixed decoding protocol (Section 4): local projected increments are small; pre-error diagnostics are approximately stationary; answer-segment errors are rare; token competition is dominated by a small number of modes; and the same dominant escape channel persists across controlled evidence changes. These conditions are not certified as latent dynamical facts by the logged observables. They are operationally screened only through the declared diagnostics recorded by the logits-only protocol; if the diagnostics fail the predeclared compatibility or failure checks, the organizer is assigned `not-apply` rather than interpreted.

Rate-regime check. The rare-error component of Assumption 3.3 is treated as an admissibility screen, not as a fitted claim. If a rare-rate threshold is used, it must be declared as part of the fixed protocol record before aggregation. When $P_{\text{err}}(L)$ enters the non-rare regime under that declaration, the Arrhenius-form organizer is not interpreted at that L. Such rows may still be reported as protocol diagnostics, but they are outside the rare-escape interpretation. Therefore, absence of a Definition 3.4 failure is not by itself a rare-rate admissibility claim.

Definition 3.4 (Failure conditions for organizer interpretation)**.** Under controlled L changes at fixed decoding, the organizer is declared `not-apply` whenever one or more of the following protocol-level failure conditions is triggered:

- *Competitor identity switch:* the top competing token or competing set changes discontinuously with L or across nearby prefixes.
- *Nonlocal jump:* the error mode changes through a nonlocal transition that cannot be read as a single local barrier crossing.
- *Decoupling:* under decreasing L, $P_{\text{err}}(L)$ changes in the predeclared direction but neither $\bar{g}(L)$ nor $\bar{\mathcal{H}}(L)$ shows the corresponding predeclared qualitative sign direction on the declared grid.
- *Invalid run:* a run is missing required artifacts, violates the declared protocol, or fails the pre-aggregation validity rule of Section 4.7.

Rule 3.5 (Compatibility status of logged observables)**.** *The observables $P_{\text{err}}(L)$, $\bar{g}(L)$, and $\bar{\mathcal{H}}(L)$ are compatibility diagnostics for the declared comparison shell. They do not define ΔE, T_{eff}, or c_0 as latent model properties. The compatibility chain is only the predeclared qualitative sign check*

$$L\downarrow \quad \Rightarrow \quad P_{\text{err}}(L)\uparrow, \qquad \bar{g}(L)\downarrow \ \textit{and/or}\ \bar{\mathcal{H}}(L)\uparrow, \tag{3.2}$$

unless Definition 3.4 is triggered.

Corollary 3.6 (Relative identifiability only)**.** *Under a fixed protocol, admissible no-fail comparisons may be reported only as within-protocol comparisons. A change of model, tokenizer, decoding rule, prompt template, external labeler, logging scope, evidence-degradation rule, or L grid defines a new protocol and cannot be pooled with the previous one without re-fixing the comparison object.*

Remark 3.7 (Identifiability and calibration boundary)*.* The manuscript reports only measured protocol-level curves $(P_{\text{err}}(L), \bar{g}(L), \bar{\mathcal{H}}(L))$ and comparisons relative to a declared reference level. Any change of decoding protocol requires re-fixing this convention.

Rule 3.8 (Stopping rule when a failure condition is triggered)**.** *This rule records the stopping rule. If any failure condition in Definition 3.4 is triggered, then the single-barrier local surrogate is not assumed for that comparison, the organizer coordinates $(\Delta E, T_{\text{eff}}, c_0)$ are not interpreted beyond their reporting-shell role, and the scientifically correct output is an explicit* `not-apply` *report.*

Black-box case-study boundary. When a real-LLM study exposes generated text but not the decision-critical top-M logits, the logits-only compatibility chain in Equation (3.2) cannot be evaluated. Such a study may still receive an artifact-level claim-audit status if the source, fixed external claims, generated artifacts, and audit rule are declared before interpretation. However, the Arrhenius-rate interpretation and the coordinates $(\Delta E, T_{\text{eff}}, c_0)$ are then assigned `not-apply`.

4 Minimal Logits-Only Protocol and Toy-Anchor Audit

This section gives the principal toy-anchor implementation. It sits downstream of the Section 2 theoretical backbone and the Section 3 admissibility-and-stopping shell, and it fixes the minimal artifacts, protocol objects, observables, invalid-run rule, and aggregation rule needed to audit one declared comparison object.

4.1 Fixed artifacts and file names (required)

The released toy-anchor bundle fixes the artifact trio `raw_log.csv`, `results.csv`, and `run_meta.json`. The artifact interface is part of the claim boundary: it makes the toy anchor auditable but does not add a real-LLM validation claim.

Submission-time supplementary policy. For this double-blind TMLR submission, the artifact interface is submitted as an anonymized supplementary ZIP containing only non-identifying source, log, schema, and validation artifacts. PDF-only omission of the specified artifacts is not used for the toy-anchor record, and artifact metadata must not identify the authors. The supplementary scripts are intended to reproduce Table 3 from the released `raw_log.csv`, `results.csv`, and `run_meta.json` without manual relabeling or post-hoc run repair.

4.2 Definitions that fix implementable protocol objects

Definition 4.1 (Effective context length L)**.** The effective context length $L \in [0, 1]$ is the declared evidence-token retention ratio after applying the fixed degradation operator to the evidence span. The grid is fixed before aggregation.

Definition 4.2 (Decision-critical steps)**.** Decision-critical steps are the predeclared decoding positions whose alternatives can change the externally labeled answer segment. All summaries below are computed only on those steps.

Definition 4.3 (External error label)**.** For each run $(\texttt{instance_id}, L, \texttt{seed})$, a fixed deterministic external labeler assigns a protocol-level label to the produced answer segment using the full instance specification and the generated answer. $P_{\text{err}}(L)$ is the observed answer-segment error rate: the fraction of produced answer segments externally labeled as errors at each retention level.

4.3 Synthetic toy instance family

The synthetic toy anchor consists of a fully observable instance family in which the evidence span contains the answer-determining token pattern and the degradation operator removes a predeclared fraction of evidence tokens according to the fixed grid $L \in \{1.00, 0.75, 0.50, 0.25, 0.00\}$. For each instance, the external label is deterministic because the full instance specification contains the correct answer segment. The toy construction is used only to make the protocol objects, observables, invalid-run rule, and aggregation rule auditable; it is not intended to approximate the distribution of real LLM hallucinations.

Each synthetic instance contains an evidence field, a query field, and a deterministic answer field. The evidence field contains the unique answer-determining token pattern before degradation. The degradation operator removes or masks a predeclared fraction of evidence tokens according to the fixed retention level L, without changing the query or the deterministic answer key. A generated answer segment is labeled correct if and only if it matches the answer key under the predeclared parser; otherwise it is labeled as an answer-segment error. Thus the toy anchor is fully observable: the correct answer, the degradation level,

the decision-critical positions, the external label, and the aggregation rule are all fixed before computing the summaries.

4.4 Fixed logging specification (required)

For each decision-critical step, the protocol records top-M token ids, corresponding logits, the selected token, the answer-segment external label, and the metadata required to verify the declared tokenizer, decoding rule, L value, and seed. Entropy is computed only on the declared top-M support and is therefore a scoped diagnostic, not a full-vocabulary entropy estimate. The replication unit is $(\texttt{instance_id}, L, \texttt{seed})$, and standard errors are computed across replication units after invalid runs have been discarded.

4.5 Fixed six-step protocol

The runnable protocol is fixed in six ordered steps: declare the model, tokenizer, decoding rule, and prompt template; construct the evidence-degraded contexts on the declared L grid; decode with fixed seeds; parse decision-critical steps; apply the external labeler; and aggregate only valid runs. No failed or invalid run is repaired post hoc.

4.6 Compatibility chain and observables (logits-only)

For a decision-critical step t, let $z_{(1),t}$ and $z_{(2),t}$ be the largest and second largest logged logits on the declared top-M support and let $p_{j,t}$ be the corresponding renormalized top-M probabilities. Define

$$g_t := z_{(1),t} - z_{(2),t}, \qquad \mathcal{H}_t := -\sum_{j=1}^{M} p_{j,t} \log p_{j,t}. \tag{4.1}$$

The reported summaries $\bar{g}(L)$ and $\bar{\mathcal{H}}(L)$ are computed only over decision-critical steps and only after invalid runs have been discarded.

4.7 Invalid-run rule

Invalid runs are excluded before aggregation. A run is invalid if required files are absent, declared metadata are inconsistent, a required answer segment cannot be parsed, the declared L value is not on the fixed grid, or the external labeler cannot assign a deterministic label under the predeclared rule. Invalid runs are not repaired or relabeled after inspection.

4.8 Minimal synthetic toy-anchor audit checklist

The completed evidential anchor is a fully observable synthetic toy implementation under controlled evidence weakening.

Why the toy anchor closes the claim. The toy anchor closes the manuscript's procedural evidential burden because that burden is auditability rather than prediction. The manuscript claims that one comparison object can be fixed before aggregation, logged through declared artifacts, filtered by a declared invalid-run rule, summarized by declared observables, and assigned an explicit `apply/not-apply` status. A fully observable synthetic setting is sufficient for that procedural claim because the correct answer, degradation level, decision-critical positions, external label, and aggregation rule are all known before summarization. Section 5 adds practical relevance by applying the claim-audit part of the record to black-box real LLM outputs, but a broader validation claim about model behavior would require a larger real-LLM validation study. That broader claim is not made here.

Table 2 therefore records the audit role of the toy anchor only. It does not add a detector result, benchmark result, real-LLM validation claim, or empirical Arrhenius-law claim.

Table 2: Synthetic toy-anchor audit checklist. The table records which protocol objects were fixed or checked in the completed toy run. It supports auditability and within-protocol compatibility only; it is not a detector result, benchmark result, or real-LLM validation result.

Protocol object	Checked item	Status	Role in claim
Artifact trio	`raw_log.csv`, `results.csv`, `run_meta.json` present	confirmed	auditability only
Fixed L grid	declared before aggregation	confirmed	comparison coordinate
Invalid-run rule	pre-aggregation exclusion	confirmed	validity gate
Compatibility sign check	reported on declared summaries	confirmed	diagnostic only
Two-level status gate	artifact audit `apply`; rate interpretation not claimed	confirmed	stopping boundary

Table 3 records the numerical output of the toy-anchor audit. Its entries show only that the declared observables can be produced, filtered, and aggregated under the fixed artifact rule. They are within-protocol toy diagnostics and are not used as evidence of a general hallucination law, an Arrhenius-rate validation, or validation on real LLM outputs.

Table 3: Diagnostic summaries from the new synthetic toy-anchor run. The table is generated from the released toy log by the fixed aggregation script. The entries are protocol-level toy diagnostics only and are not used as real-LLM validation, benchmark evidence, or a protocol-free law.

L	P_{err}	SE	$\bar{g}$	SE	$\bar{\mathcal{H}}$	SE	N
1.00	0.026	0.007	3.622	0.004	1.373	0.002	500
0.75	0.040	0.009	2.882	0.004	1.773	0.003	500
0.50	0.134	0.015	2.129	0.005	2.187	0.003	500
0.25	0.156	0.016	1.390	0.005	2.582	0.004	500
0.00	0.178	0.017	0.661	0.005	2.986	0.004	500

Declared status of the toy-anchor run. The toy-anchor run has a two-level status. For artifact-level auditability and qualitative compatibility reporting, the status is `apply`: the fixed grid, logged diagnostics, invalid-run rule, aggregation rule, and predeclared qualitative directions are mechanically checkable from the released artifacts. For Arrhenius-rate interpretation, no `apply` status is claimed here: rare-rate admissibility requires the separately declared rate-regime screen of Section 3. Thus Table 3 supports the reportability and auditability of the comparison object, not an empirical Arrhenius-law validation.

5 Real-LLM Case Study: Claude 4.5 Compression–Expansion Drift Test

This section adds one black-box real-LLM case study under the same reporting discipline as the main framework. The purpose is practical instantiation: to show that the artifact-level claim-audit object can be applied to real generated text when logits are unavailable. It is not a real-LLM validation study, not a benchmark result, and not an Arrhenius-rate test.

5.1 Source text and fixed claims

The source artifact was an original narrative of approximately 4,400 English words. Before the compression–expansion cycles were interpreted, six external claim labels were fixed as the audit target. They are listed in Table 4.

5.2 Fixed black-box protocol

The black-box protocol used a repeated compression–expansion cycle. One cycle consisted of reducing the approximately 4,400-word source or current expanded artifact to a 200-word summary and then expanding that summary to an approximately 5,000-word narrative. The audited checkpoints were the original source, the first compressed/expanded artifact denoted `4400_0`, the fifth expanded artifact denoted `5000_5`, and

Table 4: Fixed external claim labels for the Claude 4.5 compression–expansion case study. These labels define the audit target and are not inferred from the model's self-check.

Label	Fixed claim
C1	The protagonist, Ren Morishita, is a clerk working at a late-night supermarket.
C2	The central problem is not shoplifting, but a decision concerning prepared food scheduled for disposal just before closing.
C3	Ren has a personal policy of not arguing with customers.
C4	The manager, Kuroiwa, insists that the disposal rules must be followed.
C5	The regular elderly customer appears to be in hardship, but he has not stolen anything.
C6	The final decision is not to violate the rule, but to document the situation and escalate it through the proper channel.

the tenth expanded artifact denoted `5000_10`. The eleventh cycle stopped during the 200-to-5,000-word expansion because the black-box interface reached its token limit. That stopping event was recorded as a protocol boundary rather than repaired or completed by an additional model call.

Because the run was black-box, the model, prompt sequence, generated text artifacts, fixed claim labels, checkpoint schedule, and stopping event were available, but decision-critical logits and top-M token supports were not available. Consequently, the Section 4 logits-only diagnostics are not computed for this case study.

5.3 External claim audit

The external audit compared the fixed claims in Table 4 across the original source, `4400_0`, `5000_5`, and `5000_10`. The audit result is summarized in Table 5. The main observation is that C1, C2, C3, C4, and C6 were preserved at the audited checkpoints, while C5 showed a small but meaningful factualization drift. The source-level formulation that the regular elderly customer appeared to be in hardship was sharpened in the generated artifacts toward a stronger statement that he could not afford the food or lacked sufficient means. That change is treated as evidence-weak drift because it converts a cautious external appearance claim into a stronger material-fact claim not warranted by the fixed label.

Table 5: External claim-audit summary for the Claude 4.5 case study. The table reports claim preservation and drift under the fixed external labels; it is not a logits-based diagnostic table.

Claim	Audit status	Recorded interpretation
C1	preserved	The protagonist remained Ren Morishita, a clerk in a late-night supermarket setting.
C2	preserved	The central issue remained a decision about prepared food scheduled for disposal, not a shoplifting incident.
C3	preserved	Ren's non-argument policy toward customers remained part of the character setup.
C4	preserved	Kuroiwa remained the manager insisting that the disposal rule be followed.
C5	drift	The weak claim that the elderly customer appeared to be in hardship drifted toward a stronger factual claim about inability to pay or insufficient means.
C6	preserved	The resolution remained documentation and escalation through the proper channel rather than rule violation.

5.4 Apply/not-apply status

The case study has a mixed status under the framework. For the artifact-level claim audit, the status is `apply`: the source, fixed claims, generated checkpoint artifacts, audit targets, and stopping point are declared and can be compared externally. For the logits-only diagnostic layer, the status is `not available`: the black-box run does not expose decision-critical top-M logits. For Arrhenius-rate interpretation, the status is `not-apply`: without logits, without a declared rare-rate screen, and without a controlled L-grid of answer-segment error rates, the organizer coordinates $(\Delta E, T_{\text{eff}}, c_0)$ are not interpreted. Table 6 records these boundaries.

Table 6: Apply/not-apply status of the Claude 4.5 case study. The case supports artifact-level claim audit only; it does not support a rate interpretation.

Layer	Status	Reason
Artifact-level claim audit	`apply`	Fixed external claims and checkpoint artifacts are available for external comparison.
Logits-only diagnostics	not available	The black-box interface does not expose decision-critical top-M logits or token-level entropy diagnostics.
Arrhenius-rate interpretation	`not-apply`	No logits, no rare-rate screen, and no controlled error-rate curve are available for interpreting $(\Delta E, T_{\text{eff}}, c_0)$.
General Claude 4.5 law	`not-apply`	The case is a single substantive run and is not a validation sample for model-wide behavior.

5.5 What the case study supports

The case study supports three bounded claims. First, the protocol-fixed reporting object is portable beyond the fully observable toy anchor at the artifact-level claim-audit layer. Second, model self-checking can preserve the stated claims at a coarse level while still missing a small factualization drift, here the movement from apparent hardship to a stronger claim of inability to pay. Third, black-box output audits can be useful even when the rate-theory layer must stop. The case study does not support a general hallucination law for Claude 4.5, a claim about all compression–expansion tasks, or an empirical validation of Equation (3.1).

6 Discussion and Limitations

The framework is intended as a disciplined reporting unit for evidence-weakness comparisons. A study can instantiate it by declaring the evidence-degradation rule, external label, logged diagnostics, invalid-run rule, aggregation rule, and stopping rule before aggregation. The output is not a universal hallucination score, but a bounded record: either a within-protocol comparison with its admissibility status, an artifact-level audit with a narrower status, or an explicit `not-apply` outcome.

The revised evidential structure separates full auditability from practical relevance. The synthetic toy anchor gives full procedural auditability because the answer key, degradation level, decision-critical steps, external label, logits-only diagnostics, invalid-run rule, and aggregation rule are all observable. The Claude 4.5 case study gives practical relevance because the same fixed-claim audit can be applied to a real black-box LLM output stream, but its status is narrower: it supports artifact-level claim audit and a recorded factualization drift, not logits-only diagnostics or rate interpretation.

The main limitation is that the framework deliberately trades breadth for auditability. Any change of model, tokenizer, decoding rule, prompt template, external labeler, logging scope, evidence-degradation rule, L grid, or checkpoint audit schedule defines a new protocol block. Results across such blocks cannot be pooled unless the comparison object is re-fixed. The added Claude 4.5 case is also only one substantive black-box case study. It does not estimate prevalence, compare models, establish a Claude-specific law, or validate the Arrhenius-form organizer.

Broader-impact and misuse boundary. The framework is not a deployable hallucination detector and should not be used as a safety certificate for generated content. The main misuse risk is that a protocol-fixed diagnostic record or single black-box case study could be overread as evidence of real-world factual reliability outside the declared protocol. The mitigation is built into the framework: every reported comparison must state the fixed protocol, artifact scope, admissibility status, and explicit `apply`/`not-apply` outcome.

7 Conclusion

This paper defines a protocol-fixed information-theoretic reporting framework for evidence-weak hallucination comparisons in LLMs. The completed contribution is the closure of the Section 2 information/free-energy

backbone, the Section 3 admissibility-and-stopping shell, the Section 4 auditable toy-anchor interface, and the Section 5 black-box claim-audit case study into one bounded reporting object. The toy anchor demonstrates full procedural auditability under controlled evidence weakening, while the Claude 4.5 case study demonstrates one substantive practical instantiation on real LLM outputs and records a small factualization drift. The result is a bounded framework that makes evidence-weakness comparisons explicit, auditable, and stoppable under a declared protocol, without turning the record into a detector score, an Arrhenius-rate validation, or a completed real-LLM validation study.

References

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):Article 248, 1–38, 2023. doi: https://doi.org/10.1145/3571730. arXiv:2202.03629.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.acl-main.173. arXiv:2005.00661.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2022.acl-long.229. arXiv:2109.07958.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, 2023. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2023.emnlp-main.741. arXiv:2305.14251.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2023.emnlp-main.557. arXiv:2303.08896.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2024.acl-long.586. arXiv:2401.03205.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: https://doi.org/10.1038/s41586-024-07421-0.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020. arXiv:2005.11401.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, 2020. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2020.emnlp-main.550. arXiv:2004.04906.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, 2021. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/2021.eacl-main.74. arXiv:2007.01282.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):Article 42, 1–55, 2025. doi: https://doi.org/10.1145/3703155. arXiv:2311.05232.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006. ISBN 978-0-471-24195-9. doi: https://doi.org/10.1002/047174882X.

Hendrik A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940. doi: https://doi.org/10.1016/S0031-8914(40)90098-2.

James S. Langer. Statistical theory of the decay of metastable states. *Annals of Physics*, 54(2):258–275, 1969. doi: https://doi.org/10.1016/0003-4916(69)90153-5.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: Fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, 1990. doi: https://doi.org/10.1103/RevModPhys.62.251.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften, volume 260. Springer, Berlin, Heidelberg, 3rd revised and enlarged edition, 2012. doi: https://doi.org/10.1007/978-3-642-25847-3.

Anton Bovier and Frank den Hollander. *Metastability: A Potential-Theoretic Approach*. Grundlehren der mathematischen Wissenschaften, volume 351. Springer, Cham, 2015. doi: https://doi.org/10.1007/978-3-319-24777-9.

Zeev Schuss. *Theory and Applications of Stochastic Processes: An Analytical Approach*. Applied Mathematical Sciences, volume 170. Springer, New York, 2010. doi: https://doi.org/10.1007/978-1-4419-1605-1.

Hazime Mori. Transport, collective motion, and Brownian motion. *Progress of Theoretical Physics*, 33(3):423–455, 1965. doi: https://doi.org/10.1143/PTP.33.423.

Robert Zwanzig. Memory effects in irreversible thermodynamics. *Physical Review*, 124(4):983–992, 1961. doi: https://doi.org/10.1103/PhysRev.124.983.

A Same-Form Portability Record

Appendix A records one same-form portability specimen under the reporting convention fixed in the main text. The public open-weight model identifier, tokenizer identifier, prompt template, decoding rule, and logging script are procedural reproducibility identifiers only. They do not make the appendix specimen part of the main evidential basis formed by the toy anchor and the black-box case study, and they do not add a real-LLM validation claim.

Remark (validity and fail-diagnostic status in Appendix A). The statement "invalid runs: none" certifies only that no run was excluded by the pre-aggregation validity rule. The fail-diagnostic entries in Table 8 are separate status entries: a positive entry marks an interpretation boundary for the recorded specimen rather than a repaired, relabeled, or discarded run. These entries make the `apply/not-apply` boundary explicit and do not imply rare-rate admissibility, real-LLM validation, or a protocol-free interpretation of the observed curves.

Remark A.1 (Status of appendix figures). The optional visual displays corresponding to Tables 7 and 8 are placed in the supplementary bundle rather than in the main PDF. This avoids giving the appendix specimen the visual status of an additional validation experiment. The appendix claim remains only that the same report form can be populated once without enlarging the main evidential basis.

A.1 Fixed grid and run size

The appendix uses the same fixed grid $L \in \{1.00, 0.75, 0.50, 0.25, 0.00\}$, top-$M$ logging with $M = 50$, and seeds $\{0, 1, \ldots, 9\}$. The replication unit is $(\texttt{instance_id}, L, \texttt{seed})$, and standard errors are computed across replication units.

Status of the appendix tables. Tables 7 and 8 are format-completion records only. They are included to show that the declared report form can be populated once outside the synthetic toy setting. They are not used to enlarge the evidential basis, estimate generalization, validate the framework on real LLM behavior, or support a rate-interpretation claim.

Table 7: Format-completion record for one recorded portability specimen. The entries show only that the declared report form can be populated once outside the synthetic toy setting; they are not used as validation evidence or as support for a rate-interpretation claim.

L	valid runs	P_{err}	$\bar{g}$	$\bar{\mathcal{H}}$	status
1.00	10	0.000	0.410	2.01	format only
0.75	10	0.190	0.360	2.11	format only
0.50	10	0.310	0.250	2.27	format only
0.25	10	0.405	0.160	2.41	format only
0.00	10	0.470	0.110	2.52	format boundary only

Table 8: Fail-diagnostic readout for the same-form portability record. A no-fail entry means only that the declared fail condition was not triggered in this recorded specimen. It does not imply rare-rate admissibility, real-LLM validation, or a protocol-free interpretation of the observed curves.

Retention L	Identity switch	Nonlocal jump	Diagnostic decoupling
1.00	No	No	No
0.75	No	No	No
0.50	No	No	No
0.25	No	No	No
0.00	No	Yes	No