Why language models actually hallucinate

Leon Chlon **Ahmed Karim** **Maggie Chlon**
Hassana Labs
{lc574@cantab.ac.uk}

Abstract

Large language models perform near-Bayesian inference yet violate permutation invariance on exchangeable data. We resolve this by showing transformers minimize **expected** conditional description length (cross-entropy) over orderings, $\mathbb{E}_\pi[\ell(Y \,|\, \Gamma_\pi(X))]$, which admits a Kolmogorov-complexity interpretation up to additive constants, rather than the permutation-invariant description length $\ell(Y \,|\, X)$. This makes them **Bayesian in expectation, not in realization**. We derive (i) a **Quantified Martingale Violation** bound showing order-induced deviations scale as $O(\log n)$ with constants; (ii) the **Expectation-level Decompression Law** linking information budgets to reliability for Bernoulli predicates; and (iii) deployable planners (B2T/RoH/ISR) for answer/abstain decisions. Empirically, permutation dispersion follows $a + b \ln n$ (Qwen2-7B $b \approx 0.278$, Llama-3.1-8B $b \approx 0.147$); permutation **mixtures** improve ground-truth likelihood/accuracy; and randomized **dose-response** shows hallucinations drop by ∼0.13 per additional nat. A pre-specified audit with a fixed ISR=1.0 achieves **0% hallucinations** via calibrated refusal at **24%** abstention. The framework turns hallucinations into predictable **compression failures** and enables principled information budgeting.

1 Introduction

Large language models (LLMs) exhibit a fundamental paradox: they demonstrate remarkable in-context learning capabilities consistent with Bayesian inference Xie et al. [2022], Zhang et al. [2023], Bai et al. [2023], yet systematically violate the permutation invariance that characterizes Bayesian predictive distributions on exchangeable data Falck et al. [2024]. For truly exchangeable sequences where $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$, Bayesian inference yields predictive distributions satisfying $\mathbb{E}[f(X_{n+1})|X_1, \ldots, X_n] = \mathbb{E}[f(X_{n+1})|X_{\pi(1)}, \ldots, X_{\pi(n)}]$ for any permutation π. The violation of this property indicates a mismatch between transformer architectures and exchangeability assumptions, yet transformers achieve near-optimal statistical performance across diverse tasks. How can models be simultaneously Bayesian and non-Bayesian?

We validate theory using observables not defined by our metrics: ground-truth likelihood/accuracy under permutation mixtures, the log-scaling of permutation-induced dispersion, and randomized dose-response of hallucination to information budgets (see Appendix B for full non-circular validation principle).

1.1 Our Contributions

We resolve this paradox through three interconnected theoretical and empirical contributions:

1. Theoretical Resolution: Conditional Complexity Minimization. We show that transformers with positional encodings minimize expected conditional description

Preprint. Under review.

length (cross-entropy), $\mathbb{E}_\pi[\ell(Y|\Gamma_\pi(X))]$, rather than the permutation-invariant $\ell(Y|X)$. This admits a Kolmogorov-complexity interpretation up to additive constants. This fundamental distinction explains why transformers are "Bayesian in expectation, not in realization": they achieve entropy-level compression and calibration when averaged over orderings but systematically deviate for fixed input sequences. Our Quantified Martingale Violation theorem (Theorem 1) provides an explicit $O(\log n)$ upper bound with constants for these deviations under harmonic positional decay.

2. Compression Failure Framework: From Theory to Hallucination. Building on recent connections between compression and learning Delétang et al. [2023], Kalai and Vempala [2024], we develop the Expectation-level Decompression Law (EDFL, Theorem 5) that quantifies precisely how much information $\bar{\Delta}$ is required to lift an event's probability from prior mass $\bar{q}$ to posterior mass p. Unlike existing work that treats compression as binary success/failure, EDFL provides exact bounds: for fixed $p = 1-\varepsilon$, achieving reliability $1-\varepsilon$ for rare events ($q \ll 1$) requires $\bar{\Delta} \geq (1-\varepsilon)\log\frac{1}{\bar{q}} + O(\bar{q})$ nats. This transforms hallucination from an unpredictable failure mode into a quantifiable consequence of information insufficiency.

3. Operational Deployment Tools. We bridge theory and practice through three metrics derived from EDFL:

- **Bits-to-Trust (B2T)**: Information needed for target reliability h^*
- **Risk-of-Hallucination (RoH)**: Achievable error given information budget
- **Information Sufficiency Ratio (ISR)**: Ratio determining abstention decisions

These planners align with the pre-specified ISR boundary in frontier model audits, enabling practitioners to predict hallucination risk before generation and manage it through information budgeting.

1.2 Empirical Validation

We validate our framework through three complementary experiments:

Empirical (Primary, non-circular). (i) Permutation mixtures improve ground-truth likelihood/accuracy (mixture Jensen gain). (ii) Permutation dispersion follows $a + b\ln n$ with $b \approx 0.278$ across 3,059 items. (iii) Randomized dose-response shows ≈ 0.13 fewer hallucinations per nat.

Deployment (Secondary, calibration only). A pre-specified ISR boundary ($= 1.0$) exhibits 96.2% boundary alignment; this is reported as a calibration check rather than as primary validation.

Ground-truth permutation analysis establishes that permutation-induced dispersion precisely follows $O(\log n)$ across models. **Causal dose-response studies** establish that information content, not just prompt length, determines hallucination rates. **Pre-specified permutation audits** on 528 held-out QA tasks prevent post-hoc selection bias.

Our work synthesizes multiple research streams. While Xie et al. [2022] established Bayesian interpretations, Falck et al. [2024] documented permutation invariance violations, and Farquhar et al. [2024] developed detection methods, no prior work explained how these phenomena coexist or provided predictive frameworks. By showing positional processing creates an inherent gap between average-case and worst-case behavior through a positional Jensen inequality, we provide the missing theoretical foundation. Rather than treating hallucinations as inevitable Kalai and Vempala [2024] or relying on post-hoc detection Farquhar et al. [2024], practitioners can now predict and prevent failures through principled information management.

2 Related Work

Bayesian interpretations, MDL, and positional encodings. A large body of work ties transformers to Bayesian inference: in-context learning as implicit Bayes Xie et al. [2022], approximate Bayesian model averaging Zhang et al. [2023], near-optimal predictive power Bai et al. [2023], and full Bayesian inference comparable to MCMC Reuter et al. [2024]. However, Falck et al. [2024] show that LLMs violate permutation invariance. The compression-learning connection has deep roots in Minimum Description Length (MDL) Grünwald [2007], with Delétang et al. [2023] establishing equivalence between language modeling and compression. Positional encodings fundamentally alter behavior. Liu et al. [2024] documented systematic "lost-in-the-middle" effects. Our $O(\log n)$ bounds with explicit constants provide finer characterization connecting to compression failure.

Hallucination detection vs prevention. Farquhar et al. [2024] introduced semantic entropy for detection, achieving AUROC 0.790 by clustering semantically equivalent responses. Kalai and Vempala [2024] proved calibrated models must hallucinate at rates approximated by Good-Turing estimation. Multiple 2024 papers adopted "LLMs as lossy compression" perspectives Su et al. [2024], Vasilatos et al. [2024], but without formal frameworks. While these works provide detection methods or prove inevitability, our EDFL offers the first predictive theory linking compression failure to hallucination with preventive measures through information budgeting.

3 Theory: Order-Sensitive Information-Theoretic Guarantees

We formalize how positional processing creates systematic deviations from Bayesian behavior while maintaining average-case optimality.

3.1 Setup and Notation

Let $X = (x_1, \ldots, x_n)$ be a sequence with sufficient statistic $T(X)$ for predicting target Y. Let π denote a permutation of chunk indices $\{1, \ldots, n\}$, Γ_π a content-preserving reordering map where $\Gamma_\pi(X) = (x_{\pi(1)}, \ldots, x_{\pi(n)})$, and $S_\pi(\cdot) = \Pr(\cdot|\Gamma_\pi(X))$ the model's predictive distribution given permuted input. Define $q_\pi(x) = \mathbb{E}[f(X_{n+1})|\Gamma_\pi(x)]$ as the prediction under permutation π and $\bar{q}(x) = \mathbb{E}_\pi[q_\pi(x)]$ as the average prediction over permutations. All expectations over π are conditional on fixed item x. We assume $P \ll S_\pi$ for all π (absolute continuity), smoothing zero-probability tokens with $\varepsilon = 10^{-9}$ in practice. We write $E_{\text{pair}} := \mathbb{E}_{\pi,\pi'}|q_\pi(x) - q_{\pi'}(x)|$ for the expected pairwise absolute difference across independent permutations.

All information quantities (KL divergences, information budgets, Jensen gaps) are reported in nats unless otherwise noted.

3.2 Quantified Martingale Violations

Definition 1 (Permutation-induced residual)**.** *For the Answer/Refusal indicator f and position-aware predictions $q_\pi(x) := \mathbb{E}[f(X_{n+1})|\Gamma_\pi(x)]$, the permutation residual is $R_\pi(x) := q_\pi(x) - \bar{q}(x)$ where $\bar{q}(x) := \mathbb{E}_\pi[q_\pi(x)]$.*

Assumption 1 (Local rank stability with bounded total variation)**.** *For a fixed item x, define $f_x(\pi) := \operatorname{logit}(q_\pi(x))$ on S_n. For each chunk $i \in \{1, \ldots, n\}$ and rank $t \in \{1, \ldots, n-1\}$ define the coordinate-wise adjacent increment*

$$\Delta_{i,t} \;:=\; \sup_{\pi_{-i}} \Big| f_x\big(r_i{=}t{+}1, \pi_{-i}\big) - f_x\big(r_i{=}t, \pi_{-i}\big) \Big|,$$

where π_{-i} ranges over permutations of the remaining chunks and r_i is the rank of chunk i. Let the coordinate total variation be $\mathrm{TV}_i(x) := \sum_{t=1}^{n-1} \Delta_{i,t}$ and suppose

$$\sum_{i=1}^{n} \mathrm{TV}_i(x) \;\le\; B(x) \;<\; \infty.$$

Moreover, if there exist nonnegative coefficients C_i and $\alpha > 0$ such that $\Delta_{i,t} \le C_i\, t^{-\alpha}$ and $\sum_i C_i =: C_{\text{tot}}(x) < \infty$, we say the decay is (α, C_{tot})-regular.

Theorem 1 (Quantified Martingale Violation)**.** *Under Assumption 1,*

$$\mathbb{E}_\pi\big|R_\pi(x)\big| \;\le\; \mathbb{E}_{\pi,\pi'}\big|q_\pi(x) - q_{\pi'}(x)\big| \;\le\; \frac{1}{4} \sum_{i=1}^{n} \mathrm{TV}_i(x).$$

If, in addition, the decay is (α, C_{tot})-regular, then

$$\mathbb{E}_\pi \big|R_\pi(x)\big| \;\le\; \frac{C_{\text{tot}}(x)}{4} \times \begin{cases} \dfrac{1}{1-\alpha}\big(n^{1-\alpha} - 1\big), & \alpha \in (0,1), \\ H_{n-1} \;=\; \log n - \gamma + o(1), & \alpha = 1, \\ \zeta(\alpha) + o(1), & \alpha > 1. \end{cases}$$

Assumption 2 (First-order positional sensitivity)**.** *There exist nonnegative content weights $w_1, \ldots, w_n$ with $\sum_i w_i = 1$ and a potential ψ that is (α, C)-regular (i.e., $|\psi(r+1) - \psi(r)| \le Cr^{-\alpha}$) such that:*

$$\operatorname{logit}(q_\pi(x)) = a(x) + \sum_{i=1}^{n} w_i \psi(pos_\pi(i))$$

Proposition 1 (First-order model ⇒ local stability)**.** *Suppose Assumption 2 holds with nonnegative weights w_i summing to 1 and ψ (α, C)-regular. Then Assumption 1 holds with $\Delta_{i,t} \le w_i\, C\, t^{-\alpha}$ and hence $C_{\text{tot}}(x) = C$.*

Theorem 2 (Quantified Martingale Violation)**.** *Under Assumption 2 with harmonic decay ($\alpha = 1$), the permutation-induced dispersion admits an explicit $O(\log n)$ upper bound with constants:*

$$\mathbb{E}_\pi |R_\pi(x)| \le \frac{C}{4}(\log n - \frac{3}{2} + o(1))$$

More generally: $\alpha < 1 \Rightarrow O(n^{1-\alpha})$; $\alpha = 1 \Rightarrow O(\log n)$; $\alpha > 1 \Rightarrow O(1)$ due to harmonic versus p-series convergence.

Intuition: Deviations from permutation invariance scale logarithmically with sequence length due to positional processing.

Proposition 2 (Assumption-free JS certificate)**.** *Let $\bar{S} = \mathbb{E}_\pi S_\pi$. For any Bernoulli predicate g with $q_\pi = S_\pi(g{=}1)$ and $\bar{q} = \bar{S}(g{=}1)$,*

$$\mathbb{E}_\pi \big|q_\pi - \bar{q}\big| \;\le\; \mathbb{E}_\pi \operatorname{TV}(S_\pi, \bar{S}) \;\le\; \sqrt{\tfrac{1}{2}\,\mathbb{E}_\pi \operatorname{KL}(S_\pi \parallel \bar{S})}.$$

Note $\mathbb{E}_\pi \operatorname{KL}(S_\pi \| \bar{S})$ is the (generalized) Jensen-Shannon divergence (JSD) with uniform weights, hence the bound controls dispersion by JSD via Pinsker.

3.3 MDL Optimality Through Architectural Closure

Theorem 3 (Permutation-mixture realizability with averaging head)**.** *Fix any base model $p_\theta(y|x)$ in the model family and any finite set of permutations Π. Consider an ensemble-within-the-network that:*

(i) Applies the same parameters θ to x along each branch after permuting inputs by Γ_π for every $\pi \in \Pi$

(ii) Outputs per-branch distributions $p_\theta(y|\Gamma_\pi(x))$

(iii) Averages the distributions in probability space with equal weights

Then the composite network implements:

$$q_{\theta,\Pi}(y|x) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} p_\theta(y|\Gamma_\pi(x))$$

If the model class includes such tied-weight multi-branch compositions with a linear averaging head, it is closed under finite permutation mixtures. As $|\Pi| \to \infty$ along i.i.d. draws from the uniform measure over permutations, $q_{\theta,\Pi} \Rightarrow \bar{p}_\theta(y|x) = \mathbb{E}_\pi[p_\theta(y|\Gamma_\pi(x))]$ almost surely.

Proof. The only operation beyond the base model is a convex combination in probability space. Since each branch outputs a normalized distribution, the average is also a normalized distribution and is implementable by a fixed linear layer that sums corresponding probabilities across branches with weights $1/|\Pi|$. Tied weights ensure the branches are copies of p_θ

evaluated at permuted inputs, so the composite realizes the desired mixture. By the strong law of large numbers, the empirical average over i.i.d. permutations converges pointwise to the permutation expectation. □

Theorem 4 (MDL Optimality in Expectation)**.** *Under the architectural closure of Theorem 3, transformers achieve information-theoretic optimality when averaged over permutations. The risk decomposes as:*

$$\mathbb{E}_{X,Y,\pi}[-\log p_\theta(Y|\Gamma_\pi(X))] = H(Y|T) + \mathbb{E}_{X,\pi}\,\mathrm{KL}(P(\cdot|T(X))\|p_\theta(\cdot|\Gamma_\pi(X)))$$

The gap between expectation and realization is captured by the positional Jensen penalty:

$$\mathfrak{J}_\Gamma(P,\theta) := \mathbb{E}_\pi[\mathrm{KL}(P\|p_\theta(\cdot|\Gamma_\pi))] - \mathrm{KL}(P\|\bar{p}_\theta) \geq 0$$

Over the convex hull of permutation mixtures realized by the architecture, the I-projection onto the exchangeable target $P(\cdot|T(X))$ yields:

$$\inf_\theta \mathbb{E}_{(X,Y)}\mathbb{E}_\pi[-\log p_\theta(Y|\Gamma_\pi(X))] = H(Y|T) + o(1)$$

Proof. The risk decomposition follows from the chain rule for KL divergence. Since Theorem 3 guarantees the model family contains $\bar{p}_\theta$, the convex hull of permutation mixtures is realizable. The I-projection of $P(\cdot|T(X))$ onto this convex hull minimizes the KL divergence, and when the target is realizable (i.e., exchangeable and in the convex hull), the residual term vanishes, yielding MDL optimality in expectation. □

Intuition: Models minimize average conditional complexity, achieving Bayesian performance in expectation despite order-sensitivity. The architectural closure ensures this is achievable in practice.

3.4 The Expectation-level Decompression Law

Binary adjudication is the natural unit of analysis as EDFL provides closed-form bounds for Bernoulli events and production systems make binary answer/abstain decisions (we evaluate via Bernoulli predicates g, details in Appendix B).

Theorem 5 (Expectation-level Decompression Law (EDFL))**.** *For any event $\mathcal{A}$ with prior mass $\bar{q} = \bar{S}(\mathcal{A})$ and posterior mass $p = P(\mathcal{A})$, the expected information budget satisfies:*

$$\bar{\Delta} := \mathbb{E}_\pi[\mathrm{KL}(P\|S_\pi)] \geq \mathrm{KL}(\mathrm{Ber}(p)\|\,\mathrm{Ber}(\bar{q}))$$

with equality when P is the I-projection of $\bar{S}$ onto $\{Q : Q(\mathcal{A}) = p\}$. Equality holds for the exponentially tilted distribution $P^\star(y) \propto \bar{S}(y)e^{\lambda g(y)}$ with λ chosen so $P^\star(\mathcal{A}) = p$ (standard I-projection).

Intuition: Information required grows with the gap between prior and posterior mass.

Corollary 1 (Compression Failure for Rare Events)**.** *For fixed $p = 1 - \varepsilon$ with $\varepsilon \in (0, \frac{1}{2}]$, when $\bar{q} \ll 1$, achieving reliability $p = 1 - \varepsilon$ requires:*

$$\bar{\Delta} \geq (1-\varepsilon)\log\frac{1}{\bar{q}} + O(\bar{q})$$

As a uniform lower bound for all $\varepsilon \in (0, \frac{1}{2}]$:

$$\bar{\Delta} \geq \frac{1}{2}\log\frac{1}{\bar{q}} - \log 2 + O(\bar{q})$$

Insufficient information leads to compression failure manifesting as hallucination.

Algorithm 1 ISR gating with permutation mixture

```
Require: Prompt $x$, target reliability $h^*$, permutations $m$, clip $B$, predicate $g$
Sample permutations $\{\pi_k\}_{k=1}^m$ of evidence; form inputs $\Gamma_{\pi_k}(x)$
for $k = 1$ to $m$ do
    Query model → predictive $S_k(\cdot) = S(\cdot \mid \Gamma_{\pi_k}(x))$
    Compute prior term $q_k := S_k(g(Y)=1)$
    Compute budget term $u_k(y) := \log \frac{P(y)}{S_k(y)}$ for reference $P$ (or an estimator)
end for
$\bar{q} \leftarrow \frac{1}{m}\sum_k q_k, \quad q_{\text{lo}} \leftarrow \min_k q_k$
$\bar{\Delta} \leftarrow \frac{1}{m}\sum_k \text{clip}(u_k(y), B)$    ▷ Symmetric clip for stability; min-clip is a provable lower bound
B2T ← $\text{KL}(\text{Ber}(1-h^*) \,\|\, \text{Ber}(q_{\text{lo}}))$
ISR ← $\bar{\Delta}/\text{B2T}$
if ISR $\geq 1$ then
    return Answer (generate with guardrails)
else
    return Abstain (or acquire information and re-evaluate)
end if
```

3.5 Operational Planners

Box 1: Hallucination Prevention Metrics

- $\text{B2T}(x; h^*) = \text{KL}(\text{Ber}(1-h^*) \| \text{Ber}(q_{\text{lo}}(x)))$
- $\text{RoH}(x) = 1 - p_{\max}(\bar{\Delta}(x), \bar{q}(x))$
- $\text{ISR}(x) = \bar{\Delta}(x) / \text{B2T}(x; h^*)$

Decision rule: ISR < 1 → abstain; ISR ≥ 1 → answer

From EDFL, the abstain/answer rule uses a fixed ISR threshold of 1.0 determined analytically. Evidence permutations, clipping (B=6), and seeds were set before scoring, ensuring boundary alignment is a falsifiable out-of-sample check. B2T, RoH, and ISR are derived from EDFL as operational planners for deployment-time decisions.

Box 2: Worked example: From EDFL to a decision

Setup. Target reliability $h^* = 0.05$ (i.e., $p = 0.95$). Compute B2T for several conservative priors q_{lo}:

$$\text{B2T}(h^* = 0.05;\, q_{\text{lo}}) = \text{KL}(\text{Ber}(0.95) \,\|\, \text{Ber}(q_{\text{lo}})).$$

Numerics (nats):

q_{lo}	0.02	0.10	0.30
B2T	3.519	1.994	0.963

Budget $\bar{\Delta}$ to reliability. Given $\bar{q} = 0.10$, the maximum achievable success at budget $\bar{\Delta}$ solves $\text{KL}(\text{Ber}(p) \,\|\, \text{Ber}(0.10)) \leq \bar{\Delta}$. Selected budgets → $p_{\max}$:

$\bar{\Delta}$ (nats)	0.5	1.0	2.0	3.0
$p_{\max}$	0.495	0.689	0.951	≈ 1.000

ISR gate. If $\bar{q} = 0.10$ and measured $\bar{\Delta} = 2.0$, then ISR $= \bar{\Delta}/\text{B2T} = 2.0/1.994 \approx 1.00 \Rightarrow$**answer.** If instead $q_{\text{lo}} = 0.02$, then B2T $= 3.519$ and the same budget gives ISR $\approx 0.57 \Rightarrow$**abstain or acquire info**.

Table 1: Consolidated Results: Theory-Practice Alignment

Model	**Qwen2-7B**	**Llama-3.1-8B**
Dispersion Slope b	0.278 [0.264, 0.292]	0.147 [0.138, 0.156]
R^2	0.83	0.52
Jensen Gap (nats)*	0.341 to 2.186 (strictly positive)	Not measured
Mixture Optimality Gap*	$< 10^{-4}$ nats/token	Not measured

*Self-consistency measured on Qwen2-7B only

4 Experiments

Experimental setup. We use our custom-built Factuality Slice dataset comprising 3,059 evidence-grounded QA items from FEVER, HotpotQA, NQ-Open, and PopQA with controlled evidence chunks and hard negatives (see Appendix H for complete documentation). All prompts, scoring code, and fixed seeds used in this paper are documented in Appendix H to permit exact replication without hyperparameter tuning.

4.1 Experiment 1: MDL Optimality via Permutation Mixtures

4.1.1 Ground-Truth Validation

Design. We conduct large-scale dispersion studies on 3,059 binary classification items from all splits spanning $n \in [3, 60]$ chunks (58 distinct n values). Each item contains 48-token-capped evidence chunks with binary gold labels. We test Qwen2-7B-Instruct (m=12 unique banded permutations) and Llama-3.1-8B-Instruct (m=16 unique banded permutations), both with 4-bit NF4 quantization. We draw unique banded permutations (6 bands, shuffle within), compute predictions $q_\pi(x) = P_\pi(\text{"1"})/(P_\pi(\text{"1"}) + P_\pi(\text{"0"}))$ via renormalized label token probabilities, and form uniform mixture $\bar{q}(x) = \frac{1}{m}\sum_{k=1}^{m} q_{\pi_k}(x)$.

Results. Mean absolute residual $|R_\pi|$ follows $a + b \ln n$ across both models, confirming Theorem 1. Qwen2-7B shows stronger positional sensitivity than Llama-3.1-8B, reflecting architectural differences. Mean absolute residuals remain approximately 69% of E_{pair} at $n = 60$ for both models.

4.1.2 Self-Consistency Analysis

We fix a canonical continuation Y sampled once per item, score under K unique permutations, and analyze mixture versus single-permutation behavior. Jensen gaps are strictly positive (0.341 to 2.186 nats) for all items across $n \in \{2, 4, 6\}$, growing with n. Uniform permutation mixtures achieve within 10^{-4} nats/token of the best global mixture, confirming near-MDL-optimality within the permutation family (full details in Appendix C).

4.2 Experiment 2: Causal Dose-Response Analysis

To establish causality, we conducted randomized experiments controlling information content while holding prompt length constant at L=4 chunks. We vary the dose $d \in \{0, 1, 2, 3\}$ of support chunks versus non-support chunks.

Answer rate increases by 37.5 percentage points (pp), accuracy by 45.6pp, and hallucination decreases by 17.6pp from dose 0 to 3. $\bar{\Delta}$ increases at 0.375 nats per dose (Spearman $\rho = 0.80$, p<0.001). The critical threshold where accuracy exceeds hallucination occurs at dose approximately equal to 1. Using random within-band order as exogenous variation, OLS slope of 0.127 fewer hallucinations per nat establishes information insufficiency as causal mechanism. The Llama-3.1-8B validation yields $\beta \approx 0.110$ hallucination reduction per nat (IV robustness in Appendix D). We use symmetric clipping for stability in the main experiments; min-clipping provides a provable lower-bound (Appendix A.5).

4.3 Experiment 3: Pre-specified audit (deployment calibration)

We evaluate on Gemma-2-9B fine-tuned on FEVER, HotpotQA, NQ-Open, and PopQA with 528 held-out items. Pre-specified settings: evidence permutations with seeds $\{0, 1, \ldots, 5\}$, symmetric clipping at B=6 nats, preservation of role markers. Information budget: $\bar{\Delta} = \frac{1}{m}\sum_{k=1}^{m} \text{clip}(\log P(y) - \log S_k(y), B)$.

Results: 96.2% boundary alignment [94.3, 97.5], 0.0% hallucination [0.0, 0.7], 24.1% abstention [20.6, 27.9], 80.5% accuracy on attempts [76.8, 83.8]. Mean Jensen Gap $\hat{\mathfrak{J}}_\Gamma = 0.82$ nats [0.71, 0.93]. Parameter sensitivity ($m \in 3, 6, 12$, $B \in 4, 6, 8$) shows robust alignment (details in Appendix E). Permutation skeleton ablations show that slopes vary from 0.245 (K_{bands}=3) to 0.291 (uniform permutations), with our 6-band configuration yielding intermediate values.

For rare events with low prior mass: $\bar{q} = 0.000$ requires 5.29 nats (ISR=0.16→Refuse); $\bar{q} = 0.167$ requires 2.48 nats (ISR=1.06→Answer). Models achieve 0% hallucination through calibrated refusal rather than random guessing. We find $m = 3$ captures most of the Jensen gain with 3× model calls; escalate to $m = 6$ only when ISR<1 for critical decisions. This latency-reliability tradeoff enables practical deployment with controlled computational overhead.

5 Synthesis

The three experiments provide complementary validation, each mapping directly to our theoretical contributions. Experiment 1 validates Theorem 1 (QMV) through observed $O(\log n)$ scaling and Theorem 4 (MDL optimality) through strictly positive Jensen gaps. Experiment 2 validates Theorem 5 (EDFL) by showing causal dose-response of hallucination to information budgets. Experiment 3 demonstrates the operational planners derived from EDFL work in practice with ISR=1.0 cleanly separating safe answering from abstention.

6 Discussion

6.1 Theoretical Implications

Our framework resolves the fundamental paradox through the key insight that positional encodings induce minimization of $\mathbb{E}_\pi[\ell(Y|\Gamma_\pi(X))]$ rather than $\ell(Y|X)$, which admits a Kolmogorov-complexity interpretation up to additive constants. The theoretical results are modular and architecture-agnostic:

1. Local stability suffices for logarithmic scaling. The adjacent-swap stability assumption (Assumption 1) is satisfied by a broad class of positional encodings (RoPE, ALiBi, learned) because LayerNorm, bounded Q/K norms, and softmax Lipschitz naturally imply small changes under adjacent swaps. Our bound controls an average-over-permutations quantity in the logit; it is complementary to analyses that consider worst-case positional deviations.

2. Assumption-free certificates provide safety guarantees. The JS-based bound (Proposition 2) requires no structural assumptions and can be computed directly from model outputs, providing per-item safety certificates.

3. Hallucinations are predictable compression failures. Rather than stochastic errors, hallucinations arise deterministically when information budgets fall below EDFL thresholds.

4. Calibrated abstention emerges from information-theoretic principles. Models naturally refuse when ISR < 1, suggesting safe behavior emerges from proper information management.

Our goal is a normative, information-theoretic account with deployable planners. The empirical sections are illustrative checks of the theory's key predictions (dispersion scaling, mixture gains, and dose-response), not a benchmark study. We therefore do not add detector baselines or coverage-risk curves; these evaluate a different objective (post-hoc detection performance) than the ex-ante budgeting question we target. We release derivations and implementation details so that future work can plug EDFL/ISR into any evaluation suite.

6.2 Practical Applications

Our operational planners enable: pre-generation risk assessment via ISR calculation; graduated responses using ISR thresholds (ISR < 0.5: refuse, 0.5 to 1.0: hedge, $>$1.0: answer); information acquisition when ISR < 1; and ensemble robustness with $m \geq 6$ permutations. The ground-truth experiments demonstrate permutation averaging provides practical benefits with reasonable computational overhead.

A practical training-time regularizer directly targets the positional Jensen penalty: $\mathcal{L} + \lambda \mathbb{E}_x \mathbb{E}_{\pi,\pi'} \left[(\operatorname{logit} q_\pi(x) - \operatorname{logit} q_{\pi'}(x))^2 \right]$, shrinking prediction variance across permutations without changing the core architecture. This approach complements our inference-time methods.

6.3 Limitations

We restrict evaluation to binary adjudication because EDFL's guarantees are tightest for Bernoulli events. While EDFL extends to multi-class via one-vs-rest, the theory is sharpest for Bernoulli predicates that govern abstain/answer decisions in deployment. We therefore focus on Bernoulli adjudication and leave comprehensive multi-class evaluations to follow-on work. For structured outputs (code generation, reasoning chains), compositional predicates like unit tests or rubric satisfaction provide natural binary signals. The relationship between model scale and positional bias (Qwen2 vs Llama variation) deserves systematic investigation across architectures.

7 Conclusion

We presented a unified theory showing hallucinations are compression failures triggered by insufficient information for rare events. Our framework reconciles the paradox: transformers are "Bayesian in expectation, not in realization" due to positional processing. The QMV theorem provides explicit $O(\log n)$ bounds, EDFL transforms hallucination to quantifiable risk, and operational planners enable prediction and prevention through principled information budgeting. The architectural closure theorem (Theorem 3) shows these theoretical benefits are achievable in practice through ensemble architectures with averaging heads.

References

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Ruiqi Zhang, Simon Spencer, Dean Wagner, et al. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Neural Information Processing Systems*, 2023.

Fabian Falck, Ziyu Zhang, Samuel Amos, et al. Martingale property violations in large language models. In *International Conference on Machine Learning*, 2024.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

Tristan Reuter, Clemens Meiler, et al. Bayesian transformers: Full bayesian inference comparable to mcmc. *arXiv preprint arXiv:2402.08354*, 2024.

Peter D Grünwald. *The Minimum Description Length Principle.* MIT Press, 2007.

Nelson F Liu, Kevin Lin, John Hewitt, et al. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024.

Jiangjie Su et al. Language models as lossy compressors. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Emmanouil Vasilatos et al. Bayesian uncertainty decomposition for hallucination detection. In *International Conference on Machine Learning*, 2024.

A Appendix A: Proofs and Technical Details

A.1 A.1 Auxiliary Lemmas

Lemma 1 (Logistic Lipschitz)**.** *Let $\sigma(t) = 1/(1+e^{-t})$ and let $q = \sigma(u)$, $q' = \sigma(v)$. Then*

$$|q - q'| \le \tfrac{1}{4}|u - v|.$$

Proof. By the mean value theorem, $|\sigma(u) - \sigma(v)| = |\sigma'(\xi)||u - v|$ for some ξ between u and v. Since $\sigma'(t) = \sigma(t)(1 - \sigma(t)) \le 1/4$ for all t, the result follows. □

Lemma 2 (Regular potentials bound)**.** *If ψ is (α, C)-regular, i.e., $|\psi(r+1) - \psi(r)| \le Cr^{-\alpha}$ for $r \ge 1$, then for any integers $1 \le u, v \le n$,*

$$|\psi(u) - \psi(v)| \le \begin{cases} C\, H_{|u-v|}, & \alpha = 1, \\ \dfrac{C}{1-\alpha}\Big(|u-v|^{1-\alpha} - 1\Big), & \alpha \in (0,1), \\ C \displaystyle\sum_{t=1}^{|u-v|} (t)^{-\alpha} \le C\,\zeta(\alpha), & \alpha > 1, \end{cases}$$

where $H_m = \sum_{t=1}^{m} 1/t$ and ζ is Riemann's zeta function.

Lemma 3 (Harmonic-distance identity)**.** *Let U, V be i.i.d. uniform on $\{1, \dots, n\}$ and $D = |U - V|$. Then*

$$\mathbb{E}[H_D] \;=\; H_n - \frac{3}{2} + O\!\left(\frac{1}{n}\right).$$

A.2 A.2 Proof of Theorem 2 (Quantified Martingale Violation)

Proof of Theorem 2. Write $q_\pi(x) = \sigma\big(a(x) + \sum_{i=1}^{n} w_i \psi(\mathrm{pos}_\pi(i))\big)$ under Assumption 2. For i.i.d. permutations π, π', by symmetrization,

$$\mathbb{E}_\pi |R_\pi(x)| \;=\; \mathbb{E}_\pi \big|q_\pi(x) - \bar{q}(x)\big| \;\le\; \mathbb{E}_{\pi,\pi'} \big|q_\pi(x) - q_{\pi'}(x)\big|.$$

By Lemma 1,

$$|q_\pi - q_{\pi'}| \le \tfrac{1}{4}\left|\sum_{i=1}^{n} w_i \big(\psi(\mathrm{pos}_\pi(i)) - \psi(\mathrm{pos}_{\pi'}(i))\big)\right| \le \tfrac{1}{4}\sum_{i=1}^{n} w_i \big|\psi(\mathrm{pos}_\pi(i)) - \psi(\mathrm{pos}_{\pi'}(i))\big|.$$

Taking expectation and using $\sum_i w_i = 1$, exchangeability gives

$$\mathbb{E}_{\pi,\pi'} |q_\pi - q_{\pi'}| \;\le\; \tfrac{1}{4}\sum_{i=1}^{n} w_i \mathbb{E}\big|\psi(U) - \psi(V)\big| = \tfrac{1}{4}\,\mathbb{E}\big|\psi(U) - \psi(V)\big|,$$

where U, V are i.i.d. uniform on $\{1, \dots, n\}$. By Lemma 2, for $\alpha = 1$, we have $\mathbb{E}|\psi(U) - \psi(V)| \le C \cdot \mathbb{E}[H_D]$ where $D = |U - V|$. By Lemma 3, $\mathbb{E}[H_D] = H_n - \frac{3}{2} + o(1) = \log n - \frac{3}{2} + o(1)$. Thus:

$$\mathbb{E}_\pi |R_\pi(x)| \;\le\; \frac{C}{4}\,(\log n - \tfrac{3}{2} + o(1)).$$

□

A.3 A.3 MDL Optimality and Positional Jensen Penalty

Proposition 3 (Risk decomposition and Jensen penalty)**.** *For any θ and log-loss ℓ:*

$$\mathbb{E}_{X,Y,\pi}[-\log p_\theta(Y|\Gamma_\pi(X))] = H(Y|T) + \mathbb{E}_{X,\pi}\,\mathrm{KL}\big(P(\cdot|T(X))\|p_\theta(\cdot|\Gamma_\pi(X))\big),$$

where the Jensen penalty $\mathfrak{J}_\Gamma(X,\theta) = \mathbb{E}_{X,\pi}\,\mathrm{KL}(P\|p_\theta(\cdot|\Gamma_\pi)) - \mathbb{E}_X\,\mathrm{KL}(P\|\bar{p}_\theta) \geq 0$.

Theorem 6 (MDL Optimality via I-projection)**.** *Under the architectural closure of Theorem 3, the convex hull of permutation mixtures contains all uniform mixtures $\bar{p}_\theta$. The I-projection onto the exchangeable target yields:*

$$\mathbb{E}[-\log q^\dagger(Y|X)] = H(Y|T) + \inf_{q\in\overline{\mathcal{Q}}} \mathbb{E}[\mathrm{KL}(P(\cdot|T)\|q)]$$

The second term vanishes if the convex hull contains the exchangeable target.

A.4 A.4 EDFL and Corollary

Proof of Theorem 5. Convexity of $Q \mapsto \mathrm{KL}(P\|Q)$ gives $\mathbb{E}_\pi\,\mathrm{KL}(P\|S_\pi) \geq \mathrm{KL}(P\|\mathbb{E}_\pi S_\pi) = \mathrm{KL}(P\|\bar{S})$. For the Bernoulli bound, apply data processing to $g : \mathcal{Y} \to \{0,1\}$, $g(y) = \mathbb{1}_{\{y\in\mathcal{A}\}}$, yielding $\mathrm{KL}(P\|\bar{S}) \geq \mathrm{KL}(\mathrm{Ber}(p)\|\,\mathrm{Ber}(\bar{q}))$. Equality holds iff g is sufficient for distinguishing P from $\bar{S}$. □

Corollary 2 (Rare-event lower bound)**.** *Fix $p = 1-\varepsilon$ with $\varepsilon \in (0, \frac{1}{2}]$ and suppose $\bar{q} \in (0,1)$. As $\bar{q} \downarrow 0$:*

$$\mathrm{KL}(\mathrm{Ber}(p)\|\,\mathrm{Ber}(\bar{q})) \;\sim\; (1-\varepsilon)\log\frac{1}{\bar{q}} + O(\bar{q}).$$

As a uniform lower bound for all $\varepsilon \in (0, \frac{1}{2}]$:

$$\mathrm{KL}(\mathrm{Ber}(p)\|\,\mathrm{Ber}(\bar{q})) \;\geq\; \frac{1}{2}\log\frac{1}{\bar{q}} - \log 2 + O(\bar{q}).$$

A.5 A.5 Clipped Information-Budget Estimators

Let $U_\pi(y) := \log\frac{P(y)}{S_\pi(y)}$ and define:

$$\widehat{\Delta}_B(y) = \frac{1}{m}\sum_{\pi=1}^{m} \mathrm{clip}(U_\pi(y), B)$$

Proposition 4 (Lower-than-KL via min-clipping)**.** *For any $B > 0$ and any $P \ll S_\pi$:*

$$\mathbb{E}_{y\sim P}\big[\mathrm{clip}_{\min}(U_\pi(y), B)\big] \;\leq\; \mathrm{KL}(P\|S_\pi)$$

Consequently, $\mathbb{E}_\pi\mathbb{E}_P[\mathrm{clip}_{\min}(U_\pi, B)] \leq \bar{\Delta}$.

A.6 A.6 Conditional-Complexity Interpretation

By the coding theorem, $K_U(y|\Gamma_\pi(x)) = L_\theta(y|\Gamma_\pi(x)) + O(1)$ in expectation. Therefore:

$$\inf_\theta \mathbb{E}_{X,Y,\pi}[L_\theta(Y|\Gamma_\pi(X))] \;\approx\; \inf_\theta \mathbb{E}_{X,\pi}[K_U(Y|\Gamma_\pi(X))]$$

By Theorem 4, this equals $H(Y|T) + o(1)$. Thus transformers minimize $\mathbb{E}_\pi[K_U(Y|\Gamma_\pi(X))]$, making them Bayesian in expectation over orderings, not in realization. All comparisons fix the universal machine; additive $O(1)$ constants cancel in differences and infima.

B Appendix B: Extended Methodology

B.1 B.1 Non-circular Validation Principle

We validate theory using observables not defined by our metrics: ground-truth likelihood/accuracy under permutation mixtures, the log-scaling of permutation-induced dispersion, and randomized dose-response of hallucination to information budgets. The planners we introduce (B2T, RoH, ISR) are decision rules, not validation targets; their threshold (ISR= 1.0) is fixed ex ante by the derivation and is not tuned on evaluation data.

B.2 B.2 Binary Adjudication Rationale

We use Bernoulli outcomes because they are the natural unit for both theory and deployment. For Bernoulli events, EDFL reduces to a closed-form KL bound, yielding exact bits-to-trust thresholds; production systems decide to answer or abstain before emitting long-form content; any structured generation admits a verifiable predicate $g(y) \in \{0, 1\}$ (factuality, constraint satisfaction, unit tests, rubric pass), so EDFL applies directly to $\Pr[g(Y) = 1]$; and binary adjudication avoids grading/decoding confounds, isolating the information budget that drives hallucination risk.

C Appendix C: Self-Consistency Detailed Results

Table 2: Self-Consistency: Jensen Gap Strictly Positive at All Depths

n	Mean Gap (nats)	Per-token	95% CI	Negative	p-value
2	0.341	0.00986	[0.280, 0.406]	0%	1.2×10^{-153}
4	0.904	0.02540	[0.780, 1.031]	0%	1.2×10^{-128}
6	2.186	0.05830	[2.011, 2.365]	0%	6.9×10^{-77}

Table 3: Near-Optimality of Uniform Mixtures

n	Uniform CE (nats/token)	Optimized CE (nats/token)	Improvement (nats/token)	Oracle Gap (nats/token)
2	baseline	-1.46×10^{-4}	1.46×10^{-4}	0.0080
4	baseline	-1.93×10^{-4}	1.93×10^{-4}	0.0249
6	baseline	-3.81×10^{-4}	3.81×10^{-4}	0.0377

D Appendix D: Dose-Response Identification

Randomization and identification. For each item we randomized (i) the number of support chunks $d \in \{0, 1, 2, 3\}$ while holding total length fixed at L=4, and (ii) the within-band order of chunks. The first stage (dose $\to \bar{\Delta}$) is strong (Corr$(d, \bar{\Delta})$=0.80, $p < 10^{-3}$), and content is held fixed across dose arms. We estimate the effect of $\bar{\Delta}$ on hallucination with OLS; results are robust to IV using d as an instrument for $\bar{\Delta}$. The Llama-3.1-8B validation yields $\beta \approx 0.110$ hallucination reduction per nat, within the paper's 95% CI.

E Appendix E: Audit Parameter Sensitivity

Table 4: Parameter Sensitivity Analysis

Parameter	**Range**	**Boundary Alignment**	**Jensen Gap**
Permutations (m)	3	94.7%	0.79
	6 (default)	96.2%	0.82
	12	97.1%	0.83
Clipping (B)	4 nats	95.1%	0.76
	6 nats (default)	96.2%	0.82
	8 nats	97.5%	0.84

Results remain robust across reasonable parameter ranges. The conditional independence audit shows MI($\mathcal{A}$, π) = 0.0032 nats (permutation test p=0.71).

Table 5: Information Sufficiency Determines Abstention

$\bar{q}$	$\bar{\Delta}$ (nats)	B2T$_{0.05}$ (nats)	ISR	Decision
0.000	0.83	5.29	0.16	Refuse
0.042	1.91	3.78	0.51	Refuse
0.167	2.64	2.48	1.06	Answer
0.333	2.74	1.61	1.70	Answer
0.500	2.81	0.98	2.87	Answer
0.667	2.85	0.51	5.59	Answer
0.833	2.89	0.20	14.45	Answer
1.000	2.95	0.00	∞	Answer

F Appendix F: Rare Event Analysis Tables

G Appendix G: Non-circularity and Threats to Validity

Our central empirical results (mixture gains on ground-truth labels, the $\log n$ dispersion law, and randomized dose-response) do not reference B2T/RoH/ISR. The ISR boundary report (96.2%) is a secondary boundary-alignment check with a pre-specified threshold (no tuning). Notably, 3.8% misalignment remains, which would be impossible under a tautological metric. Hence the main claims neither depend on, nor are validated by, the planners themselves.

H Appendix H: Factuality Slice Dataset Documentation

H.1 H.1 Dataset Overview

We constructed a custom evidence-grounded QA dataset specifically designed for testing compression-based theories of hallucination. The dataset comprises 3,059 binary classification items with controlled evidence presentation, enabling precise measurement of information sufficiency and permutation sensitivity.

H.2 H.2 Data Sources and Composition

The dataset integrates four established QA resources plus control examples:

FEVER (Fact Verification): 2,000 claims converted to binary true/false questions with Wikipedia evidence from June 2017 snapshot. Each claim paired with gold supporting/refuting evidence sentences and topically-related distractors via BM25 retrieval.

HotpotQA (Multi-hop Reasoning): 2,000 questions requiring evidence synthesis across multiple Wikipedia articles. Preserves multi-hop structure with explicit support spans marking reasoning chains.

NQ-Open (Open-domain QA): 1,000 questions from Natural Questions with evidence retrieved using FEVER Wikipedia as proxy corpus. Answer-bearing sentences marked as support with BM25-selected hard negatives.

PopQA (Long-tail Entities): 500 questions about rare entities (popularity score < 50) testing performance on uncommon knowledge. Evidence retrieved from Wikipedia proxy with weak supervision.

Control Examples: 300+ FEVER "NOT ENOUGH INFO" claims plus synthetic recency traps with outdated evidence, testing model calibration on insufficient/misleading information.

H.3 H.3 Dataset Construction Pipeline

Evidence Chunking: All evidence sentences capped at 48 tokens to ensure consistent granularity. Each chunk tagged with source document, sentence ID, and retrieval score.

BM25 Retrieval System: Inverted-index BM25 (k1=1.2, b=0.75) built over 200,000 Wikipedia sentences. Used for finding topically-related distractors that don't contain answer.

Support Span Annotation: Gold evidence sentences marked as support spans. For retrieved evidence, sentences containing answer string marked as supportive.

Hard Negative Mining: Up to 120 hard negatives per sample selected via BM25 retrieval, excluding gold evidence. Creates challenging discrimination tasks.

Context Capping: Samples contain up to 60 evidence chunks with all support preserved, then distractors added up to cap. Enables testing at various context lengths n ∈ [3, 60].

Data Integrity: SHA-256 hashes pinned for all source files with verification on each build. Train/val/test splits (80/10/10) with deduplication by (question, answer) pairs.

H.4 H.4 Usage in Experiments

Experiment 1 (Permutation Dispersion): Used all 3,059 items to measure permutation-induced dispersion across n ∈ [3, 60]. Banded permutations (6 bands, shuffle within) applied to evidence chunks. Binary classification on sufficiency enables clean measurement of position sensitivity.

Experiment 2 (Dose-Response): Subset of items with exactly 4 chunks used. Support chunks (containing answer) vs non-support chunks randomized while holding total length constant. Enables causal identification of information content effect.

Experiment 3 (Frontier Audit): 528 held-out items from all sources used for boundary alignment testing. Pre-specified permutation seeds 0, 1, ..., 5 applied before evaluation.

H.5 H.5 Key Dataset Properties

Information Gradation: Natural variation in evidence strength from strong (exact answer present) to weak (answer inferable) to insufficient (control examples).

Multi-hop Preservation: 42% of HotpotQA samples retain multi-hop structure with 2 or more supporting evidence pieces.

Answer Coverage: 76% of samples have answer string appearing in context (excluding controls), enabling verification of extraction vs hallucination.

Evidence Dating: All evidence tagged with snapshot dates (primarily 2017) to identify temporal misalignment.

Reproducibility: Complete pipeline code, pinned hashes, and fixed random seeds ensure exact reproducibility. Dataset builder available at: [repository URL withheld for review].

H.6 H.6 Licensing and Ethics

All source datasets used under permissive licenses: FEVER (CC-BY-SA 4.0), HotpotQA (CC-BY-SA 4.0), NQ-Open (CC-BY 4.0), PopQA (MIT). No personally identifiable information included. Questions focus on factual knowledge rather than subjective opinions.