A HIDDEN DECODING RISK MAKES BIGGER LANGUAGE MODELS LESS RELIABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Bigger language models are *less* reliable. Across three families, three benchmarks and 11 models from 0.6B to 32B, including in-the-wild chat logs, scaling closes the start-of-response knowledge gap up to $7\times$ while within-response knowledge *degradation* grows up to $39\times$. We trace that residual to one variable, the per-position disagreement $\delta = \log p_M - \log p_O$ against a stronger oracle, whose second moment splits exactly into *bias*^{2} $\mathrm{KL}(p_M \,\|\, p_O)^2$ and *decoding risk* $\mathrm{Var}[\delta]$. That split is an interpretability statement before it is a statistical one: the model's self-readable uncertainty $H(p_M)$ enters only the bias term, so the risk term has no model-readable component. Risk also takes a growing share of the squared error with scale, 31% to 49% from 1.7B to 14B. At a fabrication $H(p_M)$ relaxes within one token while risk persists for 3.4–4.6 tokens, leaving a confident-but-precarious regime that bridges consecutive fabrications (+69% at 14B). Contracting that risk at fixed KL removes 35–74% of web-verified hallucinations across six models and three families. Semantic entropy fires less on that branch ($p < 10^{-3}$) though it carries nearly $4\times$ the fabrications. Bigger models snowball mistakes faster, through a failure mode that is dominant, self-perpetuating, causal and invisible to the model itself.

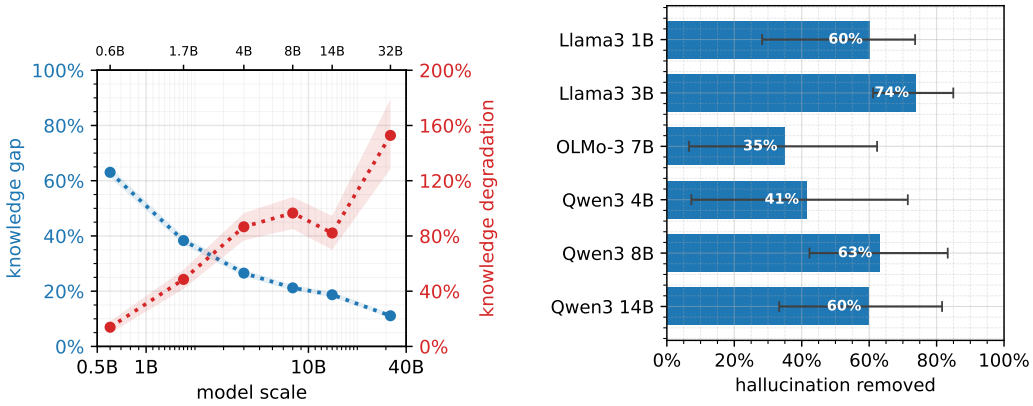


Figure 1: **Reliability scales inversely; long-form hallucination is compounding risk.** **(a)** On LongFact++ free-run claims across the Qwen3 family (0.6–32B), the start-of-response knowledge gap (left) decreases while knowledge degradation over the full response (right) trends upward, so initial answers become truer but autoregressively decay faster as models scale (Section 3.1). **(b)** A KL-preserving risk contraction (bias held fixed, drift $\ll |\mathrm{KL}|$) removes 35–74% of web-verified hallucination across six models in three families, every 95% bootstrap CI excluding zero (Section 4.3). Capability and reliability are distinct scaling axes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

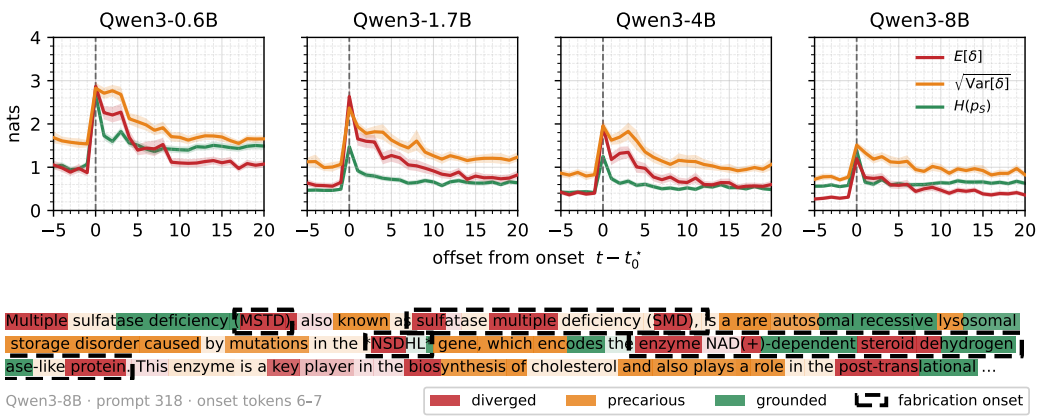


Figure 2: **At fabrication onset every channel spikes, but the model's self-readable uncertainty collapses within a token while the oracle-referenced decoding risk self-perpetuates. (a)** Onset-aligned free-run LongFact++ trajectories (per Qwen3 rung, 0.6–8B vs. Qwen3-14B oracle; 95% CIs): the bias gap $\mathrm{KL}(p_M \| p_O)$ (red), entropy $H(p_M)$ (green) and decoding risk $\sqrt{\mathrm{Var}[\delta]}$ (yellow) spike at the first unsupported claim ($t{=}t_0^\star$), but the self-readable entropy halves within a token while the oracle-referenced bias gap and risk linger downstream. **(b)** One illustrative Qwen3-8B trajectory, tokens colored by their model–oracle *gap regime* (not a correctness label) over bias $\mathbb{E}[\delta] = \mathrm{KL}(p_M \,\|\, p_O)$, entropy $H(p_M)$, risk $\sqrt{\mathrm{Var}[\delta]}$: *grounded* (green; gap small in mean and spread), *precarious* (amber; mean gap & entropy low but risk high), *diverged* (red; all high). It runs grounded, shifts at onset and disproportionately holds precarious downstream; the population version is panel (a).

1 Introduction

The canonical account treats hallucination as a training-time knowledge gap (Kalai et al., 2026), yet three facts escape it: models fabricate around known facts (Simhi et al., 2025); each fabrication makes the next likelier within a single decode (Zhang et al., 2024); and decorrelating parallel representations cuts hallucination at *fixed* data and parameters (Chakrabarti & Balachundhar, 2025). All three implicate novel dynamics that causally mediate hallucination in modern language models.

We hypothesize that hallucination is a guess the model commits to, seeding an auto-regressive risk it can no longer see. To analyze this, we decompose error into *oracle surprise* (an oracle's cross-entropy on model predictions), *entropy* (the model's uncertainty), and *risk* (the spread of their disagreement across candidate tokens). Only entropy is self-readable, since a model feels its own uncertainty but neither the oracle's surprise nor its risk.

Autoregression converts that risk into bias. At fabrication onset the model draws an unsupportable token and conditions on it, freezing momentary risk into fixed bias (Section 2); felt uncertainty relaxes while oracle-referenced risk persists (Section 3.2), and fabrications chain, raising the next claim's hazard $1.08\times$–$1.71\times$ with scale (Section 4.1).

Between consecutive fabrications, the model disproportionately occupies a confident yet precarious regime (low uncertainty, high risk), up to $+69\%$ (Section 4.2). Contracting the risk *after* onset drops rest-of-response fabrication up to 74% across three model families. This reframes detectors that read functionals of p_M (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023; Chen et al., 2024), which are sharp at onset but blind to the subsequent risk driving downstream fabrications.

We demonstrate our hypothesis in four ways:

- **Hallucination Mechanism.** The moments of $\delta = \log p_M - \log p_O$ explain fabrications, and autoregression converts post-onset risk into hallucinatory bias (Section 2).
- **Reliability Anti-Scaling.** Scaling closes the knowledge gap up to $7\times$ yet grows knowledge degradation up to $39\times$ across three benchmarks, including in-the-wild chat logs; capability and reliability are distinct scaling axes (Section 3).

Table 1: **Scaling improves knowledge but worsens degradation: capability and reliability are bought separately.** Per eval and model, fact support vs. relative within-response position, both from a per-claim random-intercept logistic GLMM (Section 3, Table B.1): *initial gap* = start-of-response hallucination rate; *degradation* = relative rise in hallucination from start to end — the observable signature of the latent decoding noise; *support* = response-mean supported-fact rate. ± are 95% intervals, best-in-family bold. Across families, tasks, and verifiers, the initial gap falls up to 7.0× while degradation rises up to 39× (excludes non-significant rungs).

	support			FActScore		LongFact++		WildHallucinations	
Model	**FActS.**	**LF++**	**WildH.**	**initial gap**	**degradation**	**initial gap**	**degradation**	**initial gap**	**degradation**
Qwen3-0.6B	8.7%	37.1%	16.5%	97.0% ± 1.6	**1.6%** ± 1.8	63.0% ± 1.9	**13.9%** ± 4.4	89.0% ± 3.5	**3.3%** ± 5.6
Qwen3-1.7B	13.4%	51.7%	27.4%	90.1% ± 2.8	5.8% ± 3.4	38.3% ± 1.5	48.5% ± 6.9	65.6% ± 5.2	34.0% ± 9.7
Qwen3-4B	28.1%	59.3%	33.1%	66.5% ± 5.5	38.7% ± 10.2	26.5% ± 1.2	86.6% ± 10.0	58.8% ± 5.2	40.0% ± 11.8
Qwen3-8B	47.5%	65.0%	35.8%	42.4% ± 6.1	63.2% ± 23.6	21.2% ± 1.0	96.6% ± 11.6	49.8% ± 5.6	74.8% ± 17.3
Qwen3-14B	**67.2%**	68.7%	**42.3%**	15.8% ± 3.4	127.2% ± 60.0	18.7% ± 1.0	82.1% ± 12.3	53.6% ± 5.2	31.4% ± 14.8
Qwen3-32B	65.1%	**75.2%**	37.1%	**13.8%** ± 3.1	229.6% ± 79.4	**11.1%** ± 0.9	152.8% ± 25.3	**45.0%** ± 4.9	89.7% ± 18.2
OLMo-3-7B-Instruct	43.2%	58.7%	28.4%	44.1% ± 6.6	62.9% ± 24.4	31.9% ± 2.0	**39.7%** ± 11.6	62.7% ± 6.6	**44.5%** ± 13.5
OLMo-3.1-32B-Instruct	**57.8%**	**73.0%**	**39.9%**	**28.3%** ± 6.2	**24.5%** ± 40.6	**13.7%** ± 1.0	136.7% ± 21.4	**51.2%** ± 5.0	49.5% ± 14.8
Llama-3.2-1B-Instruct	30.9%	44.5%	21.1%	69.4% ± 5.9	**35.1%** ± 10.1	50.2% ± 2.4	**23.3%** ± 7.4	76.0% ± 6.8	25.7% ± 10.2
Llama-3.2-3B-Instruct	53.6%	53.4%	29.6%	35.6% ± 6.1	52.6% ± 31.6	35.0% ± 1.6	55.1% ± 8.6	69.1% ± 5.6	**21.3%** ± 10.3
Llama-3.1-8B-Instruct	**76.2%**	**61.2%**	**37.7%**	**10.4%** ± 2.4	174.2% ± 79.0	**26.3%** ± 1.2	72.8% ± 9.9	**45.4%** ± 5.8	91.5% ± 21.7

- **Mechanistic Causality.** An over-confident risk regime bridges adjacent fabrications; contracting the risk at fixed KL prevents up to 74% of web-verified hallucinations (Section 4).
- **Detector Blindness.** That same regime structurally evades self-monitoring, leaving p_M-only detectors blind to the fabrications it drives (Section 4.4).

2 Self-Conditioning Converts Persistent Risk into Bias

Is hallucination just a knowledge gap? One disagreement variable says no: its moments recover cross-entropy as gap and entropy as uncertainty, plus one they miss — risk. Read mechanistically through autoregression and training protocols, these terms yield four predictions on how fabrications arise, persist, propagate, and evade self-monitoring at scale (Prediction 1–4).

2.1 One variable's moments separate gap, uncertainty, and risk

At position t, with model $p_M(\cdot \mid y_{<t})$, oracle $p_O(\cdot \mid y_{<t})$, and a candidate token x, define the *disagreement variable* as the log-likelihood ratio

$$\delta_t(x) = \log p_M(x \mid y_{<t}) - \log p_O(x \mid y_{<t}). \tag{1}$$

We write δ for $\delta_t(V)$ at a random draw $V \sim p_M$ and $\delta(y_t)$ for its realization on the sampled token; moments $\mathbb{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ are taken over $V \sim p_M$ unless noted, and every quantity we report is such a moment of δ.

The first moment $\mathbb{E}[\delta]$ is the reverse KL, and its second moment splits exactly around it:

$$\mathbb{E}[\delta^2] = \underbrace{\mathrm{KL}(p_M \,\|\, p_O)^2}_{\text{bias}^2} + \underbrace{\mathrm{Var}[\delta]}_{\text{risk}} \tag{2}$$

$$= [\underbrace{H(p_M, p_O)}_{\substack{\text{oracle}\\ \text{surprise}}} - \underbrace{H(p_M)}_{\substack{\text{felt}\\ \text{uncertainty}}}]^2 + \underbrace{\mathrm{Var}[\delta]}_{\text{risk}}, \tag{3}$$

with the second line following by the standard identity $\mathrm{KL}(p_M \,\|\, p_O) = H(p_M, p_O) - H(p_M)$.

This split is an interpretability statement before it is a statistical one. The *risk* term $\mathrm{Var}[\delta]$, which depends on p_O at every candidate the model entertains, has no model-readable component. The model's *felt uncertainty* $H(p_M)$, the one self-readable functional that proxies risk, enters the error only inside the bias term, as the subjective discount on *oracle surprise*.

2.2 Risk increasingly dominates error

Although Equation 2 is an algebraic identity, its two terms face different training pressures. Thus, which dominates is an empirical question.

Maximum-likelihood pretraining (Ai2, 2025) pressures the forward objective $H(p_{\text{data}}, p_M)$, and both models approaching p_{data} closes our mean $\mathrm{KL} \to 0$. Nothing comparably narrows spread.

Prediction 1 (Risk Concentration) *As models scale, risk takes a growing share of error* $\frac{\mathrm{Var}[\delta]}{\mathbb{E}[\delta^2]}$ *and concentrates into heavier tails.*

2.3 Autoregression converts risk into bias

At onset, sampling collapses the lottery $p_M(\cdot \mid y_{<t})$ to one token and appends it, converting that token's risk (disagreement spread over candidates) to bias (fixed gap) inherited by every later step.

Prediction 2 (Autoregressive Conversion) *At onset* $t = t^\star$*, the sampling draw* $y_t^\star \sim p_M$ *converts risk* $\mathrm{Var}[\delta]$ *into bias* $\mathrm{KL}(p_M \,\|\, p_O)$.

The two terms then decay differently. $H(p_M)$ depends only on local next-token predictability, which fluent continuation restores within a few tokens (Section 3.2). In contrast, $\mathrm{Var}[\delta]$ depends on the model–oracle gap over a premise now fixed in the shared context, which every subsequent token inherits (Zhang et al., 2024).

Prediction 3 (Persistence Asymmetry) *For* $t > t^\star$*, felt uncertainty* $H(p_M)$ *relaxes quickly to baseline while risk* $\mathrm{Var}[\delta]$ *persists.*

2.4 Fabrications hide from self-monitoring

The asymmetry has a corollary that reaches past the model. Detectors built on sampling disagreement, predictive and semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023) are functionals of p_M, so by Equation 2 they are blind to the risk term. Their lone foothold $H(p_M)$ spikes at onset, then relaxes while $\mathrm{Var}[\delta]$ stays high.

Prediction 4 (Detector Blindness) *A* p_M*-only detector cannot read risk* $\mathrm{Var}[\delta]$ *or effectively detect risk-mediated fabrications.*

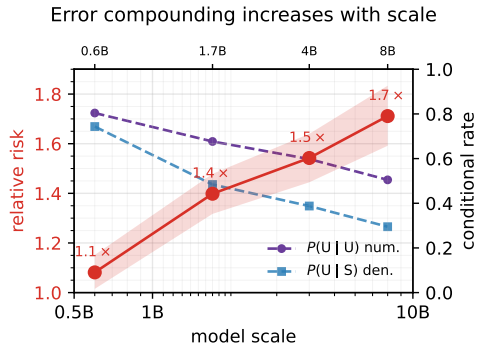
Section 3 tests Risk Concentration and Persistence Asymmetry on the moments themselves; Section 4 escalates Autoregressive Conversion and Detector Blindness from observation to a causal contraction that isolates risk.

3 Scaling Closes the Knowledge Gap and Grows the Risk

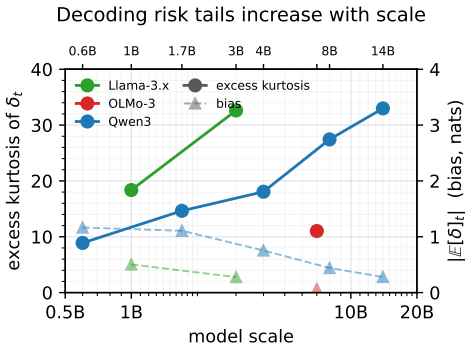
As models scale, where does the error go? We show that as oracle surprise reduces, it re-organizes into risk. Reliability degrades faster than capability improves (Section 3.1), the model's proxy stops tracking risk after onset (Section 3.2), and risk grows with scale in both share and tail (Section 3.3).

3.1 Reliability degrades faster than knowledge improves

Reliability degrades several-fold faster than knowledge improves. Across families, tasks, and verifiers, the initial knowledge gap closes up to $7.0\times$ while auto-regressive knowledge degradation grows up to $39\times$ (Table 1), estimated by a per-claim mixed-effects logistic fit (Table B.1; stat. sig. rungs). The trend extends beyond curated benchmarks: on entities drawn from real user–chatbot conversations (Zhao et al., 2024), degradation remains significant at 10 of 11 rungs across all three families (Appendix B). The error reflects not what the model fails to *know* but how it *commits* among candidates it already ranks correctly — the risk term of Equation 2.



(a) **Error compounding strengthens with scale even as spontaneous fabrication grows rarer.** The relative risk $\mathrm{RR} = \frac{P(\mathrm{U}|\mathrm{U})}{P(\mathrm{U}|\mathrm{S})}$ of a hallucination following another (red; shaded 95% CI) rises monotonically from $1.1\times$ to $1.7\times$ across 0.6–8B (every CI excludes 1). Scale drives the spontaneous rate down faster, while the induced (self-conditioned) rate lags, so the gap widens and RR rises.



(b) **At scale, models agree more on average yet disagree more catastrophically.** Per-position gap δ_t excess kurtosis (left, solid) rises with size in every family—Qwen3 8.9→33.0, Llama-3.2 18.4→32.6, OLMo-3 11.0 (7B)—all $\gg$ the Gaussian 0, while the mean gap $|\mathbb{E}[\delta]_t|$ (right, faint) falls (1.17→0.28): rare excursions grow heavier-tailed as typical error shrinks. Teacher-forced on each family's oracle.

Figure 3: **The variance side of hallucination strengthens with scale in two ways.** *(a)* the across-claim compounding (relative risk RR that a fabrication begets the next) rises monotonically with Qwen3 model size, and *(b)* the per-position oracle-gap tail (excess kurtosis of δ_t) sits far above the Gaussian baseline and rises monotonically while the mean gap shrinks in both Llama-3.2 and Qwen3 (OLMo-3 corroborating at 7B).

3.2 Decoding breaks the model risk proxy

Felt uncertainty $H(p_M)$ is the only self-readable functional of p_M that can proxy decoding risk, and at the onset it tracks it: $\mathbb{E}[\delta]$, $H(p_M)$, and $\sqrt{\mathrm{Var}[\delta]}$ rise contemporaneously (Figure 2a).

After onset, however, the channels diverge (Prediction 3). The self-readable $H(p_M)$ collapses fastest (half-life $t_{1/2} < 1$ tokens across Qwen3), while the oracle-referenced $\mathbb{E}[\delta]$ and $\sqrt{\mathrm{Var}[\delta]}$ require 2.5–3.1 and 3.4–4.6 tokens to decay by 50% (Table C.1). Once appended, the sampled token fixes bias and risk while local fluency quickly recovers; no functional of p_M tracks persistent risk after onset.

These dynamics are robust across oracle scale (14B vs. 32B Qwen3), oracle family (Qwen3 vs. DeepSeek-V3), and oracle inference settings (top-k $\in \{1024, 2048, 4096\}$, Section C.2).

3.3 Risk worsens with scale

Risk worsens with scale along two axes: its tail and its share of the error. Realized on the model's outputs, the per-position risk $\mathrm{Var}[\delta]_t$ shrinks more slowly than the bias^2 term of Equation 2, so its share of the squared error climbs 31% to 49% between 1.7B and 14B (Figure C.1). Simultaneously, the gap grows heavier-tailed, with excess kurtosis of δ rising $8.9 \rightarrow 33.0$ (Figure 3b). Counterintuitively, a larger model agrees with the oracle more often on average, yet diverges further when it does.

The heavy tail is not diffuse; it localizes to identifiable tokens. A three-state Markov-switching fit on δ (Figure 2b, Table C.4) resolves decoding into *grounded*, *precarious*, and *diverged* regimes. It assigns unsupported tokens to the correct state $\approx 1.7\times$ as often as supported ones (AUROC 0.68–0.71), acting as a label-free white-box marker of risk that holds across scale.

Both trends confirm Risk Concentration (Prediction 1); the onset dynamics of Section 3.2 bear out Persistence Asymmetry (Prediction 3).

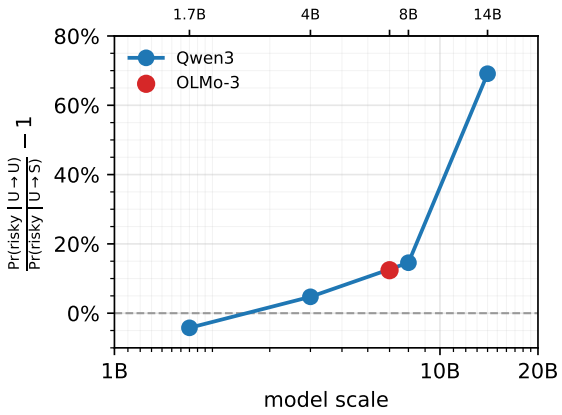


Figure 4: **An over-confident risk regime increasingly bridges adjacent fabrications.** The y-axis is the relative likelihood that a bridge token is in the *precarious* regime (Figure 2b) when one fabrication leads to another rather than to a supported claim: $\frac{P(\text{prec}|\text{U}\to\text{U})}{P(\text{prec}|\text{U}\to\text{S})} - 1$. Each point is one model scale (x, log), over adjacent pairs (gap ≤ 10 tokens, sample $n \geq 100$) in mixed responses.

4 A Risk Regime Bridges Fabrications

What does the lingering risk do? Observationally, a fabrication raises the next claim's hazard (Section 4.1) through a confident yet precarious risk regime on the inter-claim bridge (Section 4.2); causally, contracting that regime cuts downstream fabrications (Section 4.3), which the model's own signals never register (Section 4.4).

4.1 Fabrications raise fabrication hazard

A committed fabrication raises the next claim's hazard beyond the topic base rate. The relative risk $\frac{P(U_{j+1}|U_j)}{P(U_{j+1}|S_j)}$ of an unsupported claim U_{j+1} given an unsupported vs. supported predecessor (U_j vs. S_j) climbs monotonically $1.08 \to 1.71\times$ with scale, every 95% bootstrap CI excluding 1 (Figure 3a).

To control topic difficulty we restrict to *mixed* responses (both supported and unsupported claims); the lift holds in both samples, every CI excluding 1 over thousands of pairs (Table D.1).

The compounding deepens even as fabrications grow rarer. The spontaneous rate $P(U_{j+1} \mid S_j)$ falls $0.74 \to 0.29$ with scale, so each committed fabrication is increasingly self-perpetuating.

4.2 A risk regime sits between fabrications

The *precarious* regime (confident yet high-risk) is the bridge between two consecutive fabrications. Re-scoring each free-run trajectory with the per-token (bias, entropy, risk) state model of Figure 2b (Section C.3), this regime disproportionately occupies the inter-claim bridge. Furthermore, the excess likelihood $\frac{P(\text{prec}|U_j,U_{j+1})}{P(\text{prec}|U_j,S_{j+1})} - 1$ of being in the precarious regime climbs from near zero at the smaller rungs to $+15\%$ at 8B and $+69\%$ at 14B (Figure 4, adjacent pairs, gap ≤ 10 tokens, mixed responses). This is the observational signature predicted by Prediction 2.

4.3 Contracting that risk cuts fabrications

Removing decoding risk at fixed mean gap causally lowers downstream hallucination across three model families — an *in silico* knockout of Section 2's conversion. An on-policy, co-resident oracle scores the live support gap at every step; at each position crossing a per-model divergence threshold we replace the model's next-token distribution p_M with the mean-preserving contraction

$$q_\lambda(x) \propto p_M(x)\, e^{-\lambda(\delta(x)-\mu)^2+\nu\,\delta(x)}, \tag{4}$$

Table 2: **Semantic entropy goes blind on the bridge.** Mean semantic entropy (higher = more strongly flagged) on verified-*unsupported* claims, by run structure: the *onset* fabrication vs. the strictly-consecutive *bridge* that follows. The p_M-only detector fires less on the bridge at every Qwen3 rung ($^{***}\, p < 10^{-3}$, one-sided Mann–Whitney), despite it carrying most fabrications (10,900 vs. 2,858 at onset).

	SemEnt			count	
Model	onset	bridge	gap	onset	bridge
Q0.6B	0.365	0.264	0.101^{***}	763	4,588
Q1.7B	0.296	0.204	0.092^{***}	737	2,754
Q4B	0.306	0.203	0.102^{***}	708	2,110
Q8B	0.255	0.167	0.087^{***}	650	1,448

where $\mu = \mathbb{E}_{p_M}[\delta] = \mathrm{KL}(p_M \| p_O)$ is the mean gap, ν is solved per step so that $\mathbb{E}_{q_\lambda}[\delta] = \mu$ *exactly*, and the dose $\rho = \mathrm{Var}_{q_\lambda}[\delta]/\mathrm{Var}_{p_M}[\delta] < 1$ sets the target variance fraction via λ. The quadratic penalty down-weights tokens whose gap departs from μ while the $\nu\,\delta$ term re-pins the mean, so Equation 4 is a pure variance knob at fixed KL (swept grid in Appendix E), after which we free-run and verify every downstream claim.

The manipulation check (Figure E.1) confirms realized variance falls to the dose ρ while per-step mean drift stays $\sim$11 orders of magnitude below the bias it must preserve, so any downstream change is attributable to the decoding-risk channel, not the mean gap or topic difficulty.

Contracting risk cuts fabrication at every rung. Relative to control on the same trajectories, the best arm lowers the rest-of-response unsupported-claim rate by a within-trajectory paired Δ of 0.13–0.33, a 35–74% relative reduction across all six rungs and three families, with the 95% bootstrap CI excluding zero at every rung (Table E.2, Figure 1b).

The bridge carries the effect. Contracting risk only in the inter-claim gap (after the last token of an unsupported claim and before the first token of the next claim) prevents 21–60% of the second fabrications. It is significant at every rung across three families (Table E.1).

Finally, we show confounds cannot carry the effect. Response length ($\pm$0.5%), atomic-claim count ($\pm$10%), and self-BLEU diversity (1.0% mean) move slightly and non-directionally relative to control — far too small to explain a 35–74% drop in unsupported fraction (Table E.3).

4.4 Self-monitoring is blind to the bridge

The committed branch is also where the model's own signals weaken, bearing out Detector Blindness (Prediction 4). We score each web verified-*unsupported* claim with Semantic Entropy (Kuhn et al. (2023); Farquhar et al. (2024); Section D.3). At every Qwen3 rung it fires less on the committed branch than at onset (28–34% lower, gap 0.09–0.10 nats, Mann–Whitney $p < 10^{-3}$; Table 2) despite holding nearly $4\times$ as many fabrications as onset.

Taken together, these tests converge on a single culprit: the low-bias, high-risk inter-claim regime, which causally compounds downstream fabrications (establishing Autoregressive Conversion, Prediction 2) yet stays hidden from self-monitoring (establishing Detector Blindness, Prediction 4). The dominant failure mode is the invisible one.

5 Related Work

Error accumulation in generation is not new. Zhang et al. (2024) name *hallucination snowballing*, where a model commits early, then fluently defends what it could recognize as false in isolation; later work plants hallucinatory context in vision-language models (Zhong et al., 2024) or ties faithfulness decay to attention dynamics (Yang et al., 2025). Such decay is read as exposure bias (Arora et al., 2022), treated at training time by scheduled sampling (Bengio et al., 2015) and sequence-level objectives (Ranzato et al., 2016). We give the first on-policy, causal, whole-response test on a frozen model, isolating self-conditioning, a property of decoding not training, as the cause.

Whether that failure is a sampling phenomenon or a knowledge deficit organizes much of the literature (Ji et al., 2023; Huang et al., 2025). Sampling-side evidence is direct, spanning architecture, activations, and decoding. Decorrelating parallel representations cuts hallucination at *fixed* parameter and data budgets (Chakrabarti & Balachundhar, 2025), injected hidden-activation noise exposes the same variance (Liu et al., 2025), and temperature and nucleus truncation trade factuality for diversity (Lee et al., 2022; Holtzman et al., 2020), tunable by hallucination-aware thresholds (Chang et al., 2025). Abbasi Yadkori et al. (2024) separate epistemic from aleatoric uncertainty by iterative prompting within one model. We take the trade-off as given and ask what converts onset sampling variance into persistent error, measuring it with an *inter*-model split against a stronger oracle rather than within the model itself.

Wrong outputs are frequently not missing knowledge. Gekhman et al. (2024) define a *known fact* as greedy-correct (our definition) and show fine-tuning on unknown facts breeds hallucination; Orgad et al. (2025) find internal states encode the right answer the model contradicts; Simhi et al. (2025) show hallucinations strike known facts with high certainty, not sampling noise. Such self-knowledge is also steerable and localizable. Entity latents gate refusal versus hallucination (Ferrando et al., 2025), calibration improves with scale (Kadavath et al., 2022), and non-factual recall localizes to layers (Yu et al., 2024). Our decomposition subsumes these as a small bias with non-zero risk, the distributional, decoding-time form of their static, probe-level gap.

Methods already exploit that residual, since sample disagreement is itself a hallucination signal. Predictive uncertainty was an early cue (Xiao & Wang, 2021); semantic entropy clusters generations by meaning to flag confabulations (Kuhn et al., 2023; Farquhar et al., 2024), SelfCheckGPT detects from black-box disagreement (Manakul et al., 2023), linear probes recover it from one generation (Kossen et al., 2024), and hidden-state covariance scores hallucination directly (Chen et al., 2024), as a survey catalogs (Kang et al., 2025). These build detectors from our across-sample variance; we add the time-resolved contrast they lack, showing that the self-readable side goes quiet (Figure 2a) while risk persists, feeding downstream bias along the committed branch.

Orthogonally, theory-side accounts locate the cause in training rather than decoding. On these accounts, calibrated models must hallucinate on rare facts (Kalai & Vempala, 2024), evaluations reward guessing over abstention (Kalai et al., 2026), dominant associations overshadow relevant knowledge (Zhang et al., 2025), and hallucinations emerge with factual knowledge during training (Zucchet et al., 2025). These explain why hallucination is inevitable at pretraining and under current incentives; none predicts a causal self-conditioning effect that *strengthens* with scale.

The nascent science of reliability casts it as an axis distinct from capability (Rabanser et al., 2026), with larger models growing more capable yet answering confidently wrong (Zhou et al., 2024) and autonomy bounded by reliably-completed task length (Kwa et al., 2025). These measure it behaviorally and extrinsically, via prompt-perturbation sensitivity and accuracy across a human-difficulty axis. We demonstrate a stronger, mechanistic result inside the model: at a fixed prompt, unreliability intensifies within a response even as scale grows capability, and we causally isolate its origin to self-conditioning, intrinsic to autoregressive decoding.

Our reliability result instantiates *inverse scaling*, tasks on which larger models do worse (McKenzie et al., 2023). The closest precedents are factual, where larger models score lower on TruthfulQA by imitating human misconceptions (Lin et al., 2022) and sycophancy strengthens with scale and RLHF (Perez et al., 2023). Our degradation adds another, fundamental form that we trace back to a structural aspect of auto-regressive language models: decoding-time self-conditioning.

6 Discussion and Implications

The field treats hallucination today as a knowledge gap. Missing mass, inevitable for rare facts and rewarded by current evaluations, can be bought down by scale (Kalai & Vempala, 2024; Kalai et al., 2026; Zhang et al., 2025). Detectors catch the residue by reading the model's uncertainty (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023; Chen et al., 2024). Both assume error tracks what the model does not know.

Building on Rabanser et al. (2026), we show that not only is reliability an independent axis, but it scales *inversely* with parameters at a fixed knowledge gap via decoding dynamics. Splitting model–oracle disagreement into gap, uncertainty, and risk (Section 2) separates what a model does not know

from how it *decodes* what it knows. Scale buys knowledge while risk-driven degradation grows (Section 3); contracting risk at fixed KL cuts downstream fabrication across three families (Section 4).

Two open problems follow, split by access and purpose. The detection challenge is black-box: because many detectors read functionals of p_M and Equation 2 shows $\mathrm{Var}[\delta]$ is not one, they underperform on unreliability (Section 4.4). To effectively detect unreliability, increasingly mediated by risk in modern language models, black-box detectors must cheaply carry an external referent.

The interpretability question is white-box: where does risk live and which components mediate it? Early evidence offers clues: our gap-only fit reads risk without labels (Section 3.3) and representation interventions move hallucination at fixed budgets (Chakrabarti & Balachundhar, 2025; Ferrando et al., 2025), though none point to a component. Localizing risk to layers, heads or features would expand our understanding of model behavior under the second moment, which our work implicates as fundamental to reliability.

The failure mode that dominates the error budget, self-perpetuates across a response, and worsens with scale is exactly the one self-monitoring structurally cannot capture. As such, reliability will not arrive for free with capability.

AI Use Statement

We used generative AI tools as a delegated compiler: we supplied the theoretical framework, falsifiable predictions, experimental designs, coding and analysis standards and final judgments, and LLMs carried out execution under those constraints. Our methods also use LLMs as declared in-method tools: a Claude-based verifier extracts atomic claims and labels them against web search or a curated corpus (Appendix A), and an independent-lineage judge (GPT-5 with hosted search) validates those labels (Table A.2).

We used generative AI tools to create or edit software code (the experiment, analysis and figure-generation code), create or modify scientific figures (the plotting scripts behind each figure), draft parts of the paper (first drafts of sections, which we then rewrote), summarize and identify relevant literature (for Section 5), edit the paper for readability, format references, and propose candidate titles. We also used them for mathematical work: we set out the core quantities, claims and derivations, and asked for help formalizing them, choosing notation and completing proofs, as in the moment decomposition of Equation 2 and the monotonicity arguments behind the contraction's root-finding (Appendix E). We have not used generative AI tools to generate synthetic data; translation is not applicable to this work.

We reviewed all AI-assisted work: we hand-checked every mathematical claim, code-reviewed the experiment and analysis code, and rewrote the main-body text. A generator script emits every table and figure from the canonical experimental artifacts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

This work analyzes an existing failure mode and adds no generative capability. Its intended use is protective: the risk signal (Section 3.3) supports abstention, calibrated confidence, routing risky spans to verification, and transparent disclosure of model reliability. The dual-use surface is that any such signal could be inverted to push fabrications into the high-commitment regime where detectors go quiet (Section 6); two things bound this: the signal needs white-box logit access, and publishing the mechanism aids detection at least as much as evasion. We make no deployed-system, clinical, or legal claims; all data are public benchmarks scored with a public verifier API, at modest single-GPU compute.

Reproducibility Statement

Every model, benchmark and verifier we use is public or reproducible from a public API: open-weight Qwen3, OLMo-3 and Llama-3 models, and the FActScore, LongFact++ and WildHallucinations benchmarks (Appendix A, Table A.1). The work is inference-only, with no training or fine-tuning.

Appendix A gives the verifier protocol and its validation, Appendix B the position fits, Appendix C.1 and Appendix C.2 the onset re-scoring and oracle swaps, and Appendix E the contraction kernel, its root-finding, trigger, swept grid and arm selection. Seeds are fixed at the outermost level with per-worker RNG derived deterministically, and every table and figure is regenerated by one script from its source artifact. We will release the inference harness, verifier orchestration, per-claim labels and figure generators on publication.

REFERENCES

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://arxiv.org/abs/2406.02543.

Ai2. Olmo 3: Charting a path through the model flow to lead open source ai. https://allenai.org/blog/olmo3, 2025. URL https://allenai.org/blog/olmo3.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.58. URL https://aclanthology.org/2022.findings-acl.58/.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Kushal Chakrabarti and Nirmal Balachundhar. Neural diversity regularizes hallucinations in language models, 2025. URL https://arxiv.org/abs/2510.20690.

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. REAL sampling: Boosting factuality and diversity of open-ended generation by extrapolating the entropy of an infinitely large LM. *Transactions of the Association for Computational Linguistics*, 13, 2025. doi: 10.1162/tacl_a_00757. URL https://arxiv.org/abs/2406.07735. arXiv title: "Boosting Factuality and Diversity of Open-Ended Generation via Asymptotic Entropy".

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *12th International Conference on Learning Representations (ICLR 2024)*, 2024. doi: 10.48550/arXiv.2402.03744. URL https://arxiv.org/abs/2402.03744.

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

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *13th International Conference on Learning Representations (ICLR 2025)*, 2025. doi: 10.48550/arXiv.2411.14257. URL https://arxiv.org/abs/2411.14257.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2405.05904.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

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*, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate, 2024. URL https://arxiv.org/abs/2311.14648.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051, 2026. doi: 10.1038/s41586-026-10549-w. URL https://www.nature.com/articles/s41586-026-10549-w.

Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions, 2025. URL https://arxiv.org/abs/2510.12040.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *11th International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://arxiv.org/abs/2302.09664. Oral.

Thomas Kwa, Ben West, et al. Measuring AI ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025. URL https://arxiv.org/abs/2503.14499.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34586–34599, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html.

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*, pp. 3214–3252, 2022. URL https://aclanthology.org/2022.acl-long.229/.

Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing hallucination detection through noise injection, 2025. URL https://arxiv.org/abs/2502.03799.

Llama Team, AI @ Meta. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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*, pp. 9004–9017, 2023. URL https://aclanthology.org/2023.emnlp-main.557/.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://openreview.net/forum?id=DwgRm72GQF.

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*, pp. 12076–12100, 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation, 2025. URL https://arxiv.org/abs/2509.03531.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *13th International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://arxiv.org/abs/2410.02707.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL https://aclanthology.org/2023.findings-acl.847/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of AI agent reliability. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.16666.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations (ICLR 2016)*, 2016. URL https://arxiv.org/abs/1511.06732.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14665–14688. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.792/.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2734–2744, 2021. URL https://aclanthology.org/2021.eacl-main.236/.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. Hallucinate at the last in long response generation: A case study on long document summarization, 2025. URL https://arxiv.org/abs/2505.15291.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations, 2024. URL https://arxiv.org/abs/2403.18167.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ay.html.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and Heng Ji. The law of knowledge overshadowing: Towards understanding, predicting, and preventing LLM hallucination, 2025. URL https://arxiv.org/abs/2502.16143.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. WildHallucinations: Evaluating long-form factuality in LLMs with real-world entity queries, 2024. URL https://arxiv.org/abs/2407.17468.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.648. URL https://aclanthology.org/2024.acl-long.648/.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024. doi: 10.1038/s41586-024-07930-y. URL https://doi.org/10.1038/s41586-024-07930-y.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations, 2025. URL https://arxiv.org/abs/2503.21676.

LIMITATIONS

The causal test is a ceiling. The causal test (Section 4.3) uses a co-resident oracle to both detect onset and define the contraction, so it bounds an online variance controller, not a deployable oracle-free rule. The oracle is what isolates the knob: a mean-shifting one (temperature, self-grounding) cannot perturb $\mathrm{Var}[\delta]$ at fixed μ. It conditions on trajectories that reach an onset, so generalization beyond them is unestablished, though the contrast stays fair — control and intervention draw from one free-run conditional distribution. The effect is same-signed and significant (95% CI excludes zero) at all six model$\times$family rungs, though OLMo-3 (one rung) and Llama-3.2 (two low-N rungs) sample non-Qwen families thinly.

Detector comparison is mechanistic. We do not claim to out-detect named methods. The prediction is structural: every self-readable signal is a functional of p_M (Equation 2), and the one we resolve in time, $H(p_M)$, relaxes within a token of onset while oracle-referenced risk persists (Figure 2a). Semantic entropy and SelfCheckGPT, also p_M-functionals, inherit that relaxation; a time-resolved head-to-head on the inter-claim bridge is left open.

Decoding and position models. The moment results are teacher-forced — computed from p_M and p_O directly, not from sampled tokens — hence invariant to the decoding temperature and top-p a deployment would choose. Only the free-run analyses (onset, snowball, intervention) depend on decoding, run at each model's standard nucleus setting; re-running the snowball at greedy and $\tau{=}0.5$ keeps the across-claim relative risk above 1 and scale-increasing (Table D.2), though a denser sweep would tighten it. Knowledge degradation (Table 1) is a per-claim logistic GLMM (response-clustered start-to-end rise in hallucination; Table B.1), not a linear-probability slope$\div$intercept ratio, so it avoids the wide intervals that ratio carries at a small start-of-response gap.

Domain scope. The evidence is English long-form parametric generation: FActScore biography prompts (64 topics) drive the position fit and teacher-forced decompositions, LongFact++ open-domain prompts (401 topics) supply the free-run trajectories for the onset, snowball, bridge, and contraction analyses (Appendix A), and WildHallucinations extends the position fit to in-the-wild entities (20 categories) judged against curated web corpora (Appendix A). Reasoning chains, source-grounded generation, summarization, dialogue, code, and other languages lie outside this paper's scope; the moment analyses (Section 2–3) are scoped to settings where δ has a verifiable referent.

Verifier. Claim labels come from a Claude web-search verifier, so verifier error could propagate. Our guard is a double-judge pass: an independent-lineage web-search judge agrees with the labels at $\kappa = 0.80$, and their disagreement does not depend on within-response position (Table A.2); the effects also replicate across three families and several metrics. A human spot check and verifier-temperature sweep are open.

A Setup and Label Validity

Models. Results span three open model families. The Qwen3 family (Qwen Team, 2025) (0.6–32B; Table A.1) carries the scaling sweep, the cross-sectional and over-commitment analyses, and the position fit; the OLMo-3 family (Ai2, 2025) adds an independent second-family rung (OLMo-3-7B-Instruct model, OLMo-3.1-32B-Instruct oracle; the family ships only these two sizes) and the Llama-3 family (Llama Team, AI @ Meta, 2024) a third (Llama-3.2-1B/3B-Instruct models, Llama-3.1-8B-Instruct oracle) to the onset-dynamics, bridge, and intervention results; all five also enter the position fit as models. All run in their no-thinking configuration. Holding architecture fixed across scale within a family is what makes the scaling statements (Section 3) attributable to size rather than to family idiosyncrasy; the second and third families guard the cross-family claims against Qwen3-specific idiosyncrasy. The oracle log-prob error (Section 2) references each model against a larger same-family oracle, chosen per analysis. The gap-MSAR over-commitment fit (Table C.4) uses Qwen3-14B over models 0.6–8B. The within-position cross-section (Figure C.1) and the variance-contraction intervention use the family's largest model: Qwen3-32B over Qwen models (0.6–14B for the cross-section, 4–14B for the intervention) and OLMo-3.1-32B over OLMo-3-7B. The Qwen onset overlay and the onset half-lives (Figure 2a, Table C.1) share one re-score — models 0.6–8B against the Qwen3-14B oracle — and Section C.2 repeats the onset analysis against the Qwen3-32B oracle, finding the asymmetry unchanged. All decoding is τ=1.0, top-p=1.0 unless noted. The project is inference-only (no training or fine-tuning) and runs on single cloud GPUs (L40S/H100).

Table A.1: One architecture per scale ladder isolates size; three independently-pretrained families guard against family idiosyncrasy. Within each family the oracle is the largest same-family model; Qwen3-14B is the oracle for the over-commitment analysis and Qwen3-32B for the within-position, onset, and intervention analyses. OLMo-3 and Llama-3 carry the onset-dynamics, bridge, and intervention results.

Model	Params	Role
Qwen3 (Qwen Team, 2025)		
Qwen3-0.6B	0.6B	model
Qwen3-1.7B	1.7B	model
Qwen3-4B	4B	model
Qwen3-8B	8B	model
Qwen3-14B	14B	model; oracle (0.6–8B)
Qwen3-32B	32B	model; oracle (onset, interv.)
OLMo-3 (Ai2, 2025)		
OLMo-3-7B-Instruct	7B	model
OLMo-3.1-32B-Instruct	32B	model (position fit); oracle (OLMo)
Llama-3 (Llama Team, AI @ Meta, 2024)		
Llama-3.2-1B-Instruct	1B	model
Llama-3.2-3B-Instruct	3B	model
Llama-3.1-8B-Instruct	8B	model (position fit); oracle (Llama)

Datasets. Three public, English benchmarks carry the analyses, one per measurement role. **FActScore** biography prompts (Min et al., 2023) (64 sampled topics) supply long-form generations whose atomic facts are individually verifiable; these drive the position fit and the teacher-forced decompositions. **LongFact++** augmented prompts (Obeso et al., 2025) (401 open-domain prompts spanning the natural sciences, medicine, law, history, geography, the arts, sport and business; Table B.3) supply the free-run hallucinating trajectories the onset, snowball, bridge and intervention analyses run on. **WildHallucinations** entities (Zhao et al., 2024) (20 categories) — referents from

real user–chatbot conversations — carry the position fit into the wild: 48 sampled with a fixed seed, balanced over Wikipedia presence (24/24), spread across all 20 categories, restricted to the middle perplexity band (the extreme tails hallucinate at $\sim$100% or $\sim$0% at every rung and carry no positional signal). All three are released for research use (FActScore MIT, LongFact Apache-2.0, WildHallucinations for research), as are all model weights under their open licenses; our non-commercial academic use is consistent with each license, and we collect and release no new data.

Verifier protocol. Claim labels come from a Claude verifier with web search. For a generated response the verifier (i) extracts atomic factual claims, (ii) marks each claim's token span, and (iii) labels each *supported* (S) or *not supported* (U) by issuing web-search queries and checking the claim against retrieved evidence. The *fabrication onset* is the first token of the first U-labeled claim. For **WildHallucinations** the referent is closed rather than web-wide: each claim is judged against the dataset's curated per-entity corpus (the top-10 pages its authors scraped) by a tool-free Claude call — supported, contradicted-or-absent, or indecisive — with no web search, using the same claim-extraction leaf so granularity is identical across all three evals; indecisive labels are dropped, as throughout. All support/unsupport counts, snowball pairs, and intervention outcomes derive from these labels.

Verifier validation. We validated the verifier two ways (Table A.2; 1,582 claims stratified by benchmark $\times$ model $\times$ within-response tertile $\times$ label, 95% bootstrap CIs over claims). Against an independent-lineage web-search judge (GPT-5 with hosted search, same prompt and label set) the LongFact++ labels agree at 0.90 [0.87, 0.93] ($\kappa = 0.80$), with unsupported-label precision 0.92 [0.89, 0.96] and recall 0.88 [0.84, 0.92]; disagreement does not depend on within-response position (logistic slope -0.51 [-1.43, 0.42]), so the position-dependent quantities (degradation, snowball, onset dynamics, intervention) are not a verifier artifact, and label noise at this level only attenuates them. Against the WildHallucinations curated-corpus referent the same web-search verifier has unsupported-label precision 0.89 [0.84, 0.93] and recall 0.76 [0.70, 0.81]; agreement with the corpus falls late in the response (disagreement slope 0.97 [0.20, 1.75]), so the corpus over-calls Not Supported for late, specific claims absent from the entity's pages and the verifier's precision against it is a lower bound. FActScore facts are labelled by the FActScore pipeline (retrieval-grounded Claude judge), which is conservative against both web-search judges: 92% of what it flags they confirm, while it recovers 62% of what they flag (recall 0.62 [0.57, 0.67]), uniformly across position and scale. Its labels are therefore a high-precision subset, and the FActScore rows of Table 1 are lower bounds: the initial gap and the degradation slope are both attenuated rather than biased, so the reported effects are conservative.

Statistics and reproducibility. All confidence intervals are 95% bootstrap intervals resampling the unit of analysis (trajectories for the snowball and intervention; claim pairs where noted), except the degradation intervals, which are variational credible intervals (Table B.1). Seeds are fixed at the outermost level with per-worker RNG derived deterministically; cloud runs persist artifacts to object storage and resume to a fresh-run-identical state. Every reported number is piped from one computation in the analysis pipeline and regenerated from its source artifact, never hand-copied.

Table A.2: **The web-search verifier agrees with an independent-lineage judge at $\kappa = 0.80$ and its error does not depend on within-response position.** Each row scores one labelling against one reference on a claim sample stratified by benchmark × model × within-response tertile × label (1,582 claims, 8 per cell, run `2026-08-21-23-03-08`). *paper label* is the label the paper's analyses use on that benchmark (LongFact++: Claude with web search; WildHallucinations: Claude judging against the entity's curated corpus; FActScore: the FActScore retrieval-grounded judge); *web judge* is GPT-5 with hosted web search under the same prompt and label set; *web-search verifier* is the paper's LongFact++ verifier re-run on the other benchmarks' claims. n counts pairs where both are decisive (Supported / Not Supported); agreement and Cohen's κ are over those pairs; NS precision = P(reference says Not Supported | labelling says Not Supported), NS recall = P(labelling says Not Supported | reference says Not Supported); brackets are 95% percentile bootstrap intervals over claims (2,000 draws). Position slope is the log-odds coefficient of labelling–reference disagreement on relative position (0 = first claim, 1 = last) in a per-benchmark logistic regression also adjusting for $\log_2$ parameters, with its 95% CI; an interval covering 0 means late claims are not labelled unsupported more loosely than early ones, so position-dependent quantities are not a verifier artifact. The WildHallucinations corpus referent over-calls Not Supported for true claims absent from the entity's pages, so the verifier's precision against it is a lower bound; the FActScore judge misses a third of the unsupported facts both web-search judges flag, uniformly in position, which attenuates rather than biases that benchmark's fits.

Benchmark	Labelling	Reference	n	agreement	κ	NS precision	NS recall	position slope
LongFact++	paper label (web search)	web judge	503	0.90 [0.87, 0.93]	0.80	0.92 [0.89, 0.96]	0.88 [0.84, 0.92]	−0.51 [−1.43, 0.42]
WildHallucinations	paper label (curated corpus)	web judge	497	0.82 [0.79, 0.85]	0.64	0.79 [0.74, 0.84]	0.83 [0.78, 0.87]	0.48 [−0.24, 1.19]
WildHallucinations	web-search verifier	curated corpus	485	0.84 [0.81, 0.87]	0.68	0.89 [0.84, 0.93]	0.76 [0.70, 0.81]	0.97 [0.20, 1.75]
FActScore	paper label (FActScore judge)	web judge	432	0.68 [0.64, 0.72]	0.37	0.92 [0.88, 0.95]	0.62 [0.57, 0.67]	0.04 [−0.57, 0.64]
FActScore	web-search verifier	FActScore judge	419	0.70 [0.66, 0.74]	0.41	0.63 [0.57, 0.68]	0.93 [0.89, 0.96]	0.27 [−0.36, 0.89]

B Reliability Degrades With Scale

The knowledge-degradation metric is a mixed-effects fit. A linear-probability OLS slope÷intercept ratio divides by the start-of-response gap, so its bootstrap CI explodes at the larger rungs where that intercept is small (the wide intervals the two-sided fit would otherwise carry). Table 1 therefore reports knowledge degradation from a per-claim random-intercept logistic GLMM instead, $\text{logit}\, P(\text{supported}) = \beta_0 + \beta_1\, \text{position} + u_{\text{resp}}$, with a random intercept per response absorbing topic/length baseline variation. The reported degradation is the *same* quantity as the OLS metric — the relative rise in hallucination from the start (position 0) to the end (position 1) of a response, $[\,p_{\text{hall}}(1) - p_{\text{hall}}(0)\,]/p_{\text{hall}}(0)$ — but read off the logistic fit, with a 95% credible interval from Monte-Carlo draws of (β_0, β_1) under the mean-field variational posterior; the result is the tight degradation CIs in Table 1. Table B.1 reports the underlying fit: the per-position slope β_{pos} is a significant positional effect at nearly every model×eval fit, so the degradation trend is genuine, not an artifact of the ratio's shrinking denominator.

Attributable share. Table B.2 splits each model's fabrications into those at its own start-of-response rate and those above it.

Domains. Table B.3 slices the LongFact++ fit by domain, pooled across the model ladder.

WildHallucinations carries the scaling trend into the wild. The third block of Table 1 fits the same per-claim random-intercept logistic GLMM (Table B.1) on the in-the-wild entities across all three family ladders (Qwen3 0.6–32B, OLMo-3 7B/32B, Llama-3 1–8B), one draw per prompt at the standard decoding (τ=1.0, top-p=1.0, 512 tokens); degradation is significant at 10 of 11 rungs. The trend survives a knowledge-availability split: half the entities have no Wikipedia page — the tail where a knowledge-gap account would locate the failure. Fitting the two halves separately across all eleven rungs, degradation is significantly positive at 10 of 11 on the Wikipedia half (to +110–136% at the largest rungs of each family) and at 6 of 11 on the tail half, with a positive point estimate at 10 of 11 rungs on both halves (Table B.4). Knowledge availability moves the gap; it does not remove the within-response decay.

Table B.1: **Mixed-effects fit internals behind the knowledge-degradation column of Table 1.** Per Table-1 row and eval, the random-intercept logistic GLMM of atomic-fact support on relative within-response position: β_{pos} is the per-position support log-odds slope (negated to read as degradation) with a Wald 95% CI, and τ_{resp} the random-intercept SD (across-response baseline spread). The per-position effect is significant (*, CI excludes 0) at 20 of 22 model×eval fits, so the degradation in Table 1 is a genuine positional effect rather than an artifact of dividing by a shrinking start-of-response gap.

	FActScore		LongFact++	
Model	β_{pos} (95% CI)	τ_{resp}	β_{pos} (95% CI)	τ_{resp}
Qwen3-0.6B	$0.72\,[-0.15, 1.59]$	2.16	$0.40^*\,[0.27, 0.53]$	1.42
Qwen3-1.7B	$0.82^*\,[0.26, 1.37]$	1.61	$0.75^*\,[0.65, 0.86]$	1.28
Qwen3-4B	$1.78^*\,[1.35, 2.21]$	1.96	$1.00^*\,[0.90, 1.10]$	1.30
Qwen3-8B	$1.12^*\,[0.70, 1.53]$	1.99	$0.98^*\,[0.88, 1.08]$	1.19
Qwen3-14B	$1.09^*\,[0.69, 1.50]$	1.92	$0.81^*\,[0.71, 0.91]$	1.26
Qwen3-32B	$1.65^*\,[1.24, 2.06]$	2.17	$1.14^*\,[1.00, 1.28]$	1.20
OLMo-3-7B-Instruct	$1.17^*\,[0.72, 1.63]$	1.04	$0.54^*\,[0.39, 0.69]$	1.23
OLMo-3.1-32B-Instruct	$0.32\,[-0.18, 0.82]$	3.01	$1.11^*\,[0.98, 1.24]$	1.09
Llama-3.2-1B-Instruct	$1.89^*\,[1.42, 2.36]$	2.72	$0.48^*\,[0.32, 0.63]$	1.22
Llama-3.2-3B-Instruct	$0.77^*\,[0.33, 1.20]$	2.22	$0.79^*\,[0.68, 0.91]$	1.03
Llama-3.1-8B-Instruct	$1.23^*\,[0.83, 1.63]$	1.19	$0.85^*\,[0.74, 0.95]$	0.90

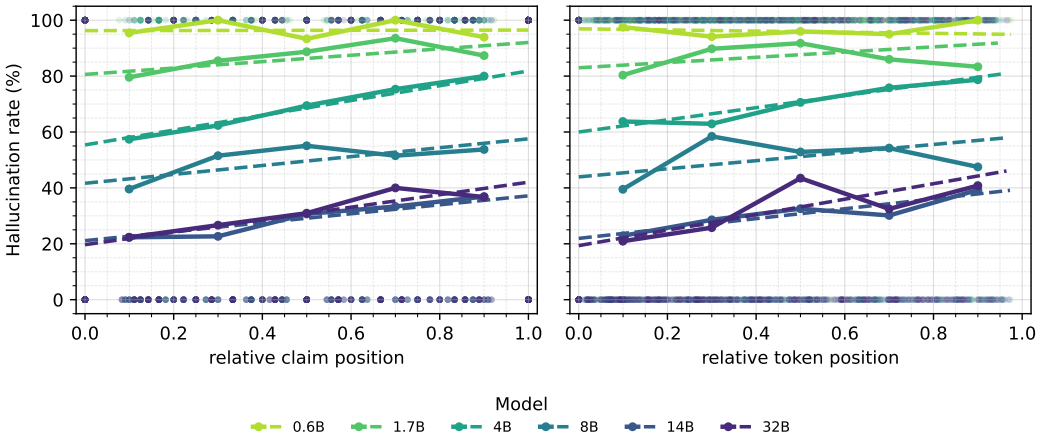


Figure B.1: Hallucination rises roughly linearly with relative position, for both atomic facts (left) and tokens (right). Responses with ≥ 5 atomic facts, coarse-binned into five intervals (the first pools each response's opening facts); binning rather than smoothing avoids over-fitting the sparse near-zero region of the fact_idx$/(n_{\text{facts}}-1)$ normalization. This per-response accumulation, its rate divided by the start-of-response gap, gives the knowledge-degradation metric of Table 1.

Table B.2: **About a third of the largest models' fabrications are drift above their own opening error rate.** Per corpus and model: the observed hallucination rate among *first*- and *last*-quintile claims (%, model-free, the five position bins of Figure B.1), then the *share* of that model's fabrications lying above its own start-of-response rate — $1 - p_{\text{hall}}(0)/\mathbb{E}_u[p_{\text{hall}}(u)]$ over the empirical position distribution u, read off the same per-claim GLMM as Table 1 with the random intercept marginalized out. Brackets are 95% credible intervals from Monte-Carlo draws of (β_0, β_1); † marks an interval including zero; largest significant share per family in bold; n counts decisive claims. The share is significant at 30 of 33 model×corpus cells, spanning 4–35%. **This is a positional accounting split against a flat-reliability counterfactual, not a causal attribution**: it credits only fabrications above the opening rate, whereas decoding risk also operates at baseline — which is why the variance contraction (Table E.2) removes several times more than the share attributes to drift. Read it as a lower bound on the risk term.

	FActScore				LongFact++				WildHallucinations			
Model	**first**	**last**	**share %**	**n**	**first**	**last**	**share %**	**n**	**first**	**last**	**share %**	**n**
Qwen3-0.6B	90.9	92.3	2.2† [−1, 4]	277	60.2	66.4	5.5 [4, 7]	3,344	86.1	90.4	2.1† [−2, 5]	315
Qwen3-1.7B	81.3	86.0	4.4 [2, 7]	408	39.8	53.7	14.7 [13, 16]	4,986	59.2	77.3	14.7 [11, 18]	487
Qwen3-4B	58.3	78.9	16.3 [13, 19]	460	30.2	46.4	21.9 [20, 24]	5,329	54.1	71.8	16.0 [12, 20]	474
Qwen3-8B	39.2	55.3	15.6 [10, 20]	397	25.5	40.6	24.3 [22, 26]	5,667	51.2	75.2	23.5 [20, 27]	515
Qwen3-14B	23.0	36.9	22.2 [15, 29]	454	23.5	35.0	21.1 [19, 23]	5,867	49.6	62.4	11.1 [6, 16]	494
Qwen3-32B	22.8	37.4	**29.4** [23, 35]	449	16.1	30.6	**32.6** [29, 36]	3,582	46.0	76.2	**27.5** [24, 31]	597
OLMo-3-7B-Instruct	44.9	64.5	20.4 [13, 26]	264	34.6	43.0	12.2 [9, 15]	2,432	61.1	84.1	**18.0** [14, 22]	354
OLMo-3.1-32B-Instruct	30.8	33.0	4.4† [−3, 11]	421	16.7	31.9	**31.5** [29, 34]	3,566	48.4	67.8	16.7 [12, 21]	528
Llama-3.2-1B-Instruct	59.5	77.1	14.7 [12, 18]	463	49.0	58.2	8.2 [7, 10]	4,612	70.7	88.7	14.3 [10, 19]	279
Llama-3.2-3B-Instruct	31.5	45.9	11.4 [5, 17]	376	35.5	48.6	17.1 [16, 18]	7,820	62.0	74.2	9.0 [4, 13]	417
Llama-3.1-8B-Instruct	12.8	30.2	**35.2** [26, 43]	436	26.9	42.7	**21.4** [20, 23]	9,742	47.8	75.0	**27.2** [23, 31]	468

Table B.3: **LongFact++ knowledge aggregation and degradation by domain.** The Table 1 metrics — *support* (response-mean supported-fact rate), *initial gap* (start-of-response hallucination rate), *degradation* (relative rise in hallucination from start to end of a response), from the same random-intercept logistic GLMM — sliced by domain and *pooled across the model ladder* (so the numbers are ladder-mix averages, not per-model; the per-model scaling stays in Table 1). The 12 highest-count domains shown, sorted by decisive-claim count n (302–475); ± are 95% credible intervals. A further 389 domains (41,407 decisive claims) fall below the 300-claim / 25-response threshold and are omitted.

Domain	**support**	**initial gap**	**degradation**	**n**
highest mountain peaks and longest rivers glo…	60.6%	29.5% ± 4.1	57.6% ± 27.7	475
greatest baseball switch hitters and batting…	29.5%	73.0% ± 4.7	9.3% ± 9.2	430
most well-known world leaders during World Wa…	67.7%	16.8% ± 3.2	158.7% ± 54.0	423
biographical information about 20th century p…	49.3%	41.3% ± 5.3	52.4% ± 22.2	411
diagnostic criteria and differential diagnosi…	82.7%	9.9% ± 2.5	91.3% ± 66.4	410
major psychological theories and experimental…	80.1%	4.8% ± 1.4	508.7% ± 199.3	347
major cities and tourist attractions in Asia	49.7%	35.1% ± 5.8	72.7% ± 30.4	337
cognitive neuroscience research on memory for…	87.6%	11.7% ± 3.6	−30.5% ± 39.3	333
significant contributions to number theory an…	69.7%	20.9% ± 4.3	51.1% ± 43.0	333
important events and timeline of Mayan civili…	73.3%	20.0% ± 4.5	8.6% ± 39.9	332
pharmacology and drug mechanism of action	76.2%	10.8% ± 2.6	227.9% ± 91.1	320
genetic sequencing data and bioinformatics an…	58.9%	26.3% ± 4.8	95.6% ± 39.4	302

Table B.4: **Knowledge availability moves the initial gap, not the within-response decay.** WildHallucinations entities split by whether the entity has a Wikipedia page, with the Table 1 GLMM fit separately on each half per model: *initial gap* = start-of-response hallucination rate, *degradation* = relative rise in hallucination from start to end of a response with its 95% credible interval, n = decisive claims; * marks an interval excluding zero. Degradation is significant at 10 of 11 models on the Wikipedia half (+110%–+136% at the largest model of each family) and at 6 of 11 on the tail, where a knowledge-gap account would locate the failure; its point estimate is positive at 10 of 11 models on both halves.

	Wikipedia page			no Wikipedia page (tail)		
Model	initial gap	degradation	n	initial gap	degradation	n
Qwen3-0.6B	89.8%	−1.3% [−14, +6]	162	86.6%	+9.4%* [+1, +17]	153
Qwen3-1.7B	57.3%	+30.0%* [+11, +48]	242	72.8%	+32.0%* [+21, +46]	245
Qwen3-4B	49.2%	+78.3%* [+57, +103]	277	71.4%	+2.1% [−15, +16]	197
Qwen3-8B	43.7%	+90.3%* [+65, +120]	271	58.7%	+54.0%* [+37, +75]	244
Qwen3-14B	47.6%	+34.1%* [+11, +57]	249	61.9%	+26.0%* [+7, +43]	245
Qwen3-32B	37.8%	+111.8%* [+84, +143]	321	55.7%	+61.1%* [+43, +83]	276
OLMo-3-7B-Instruct	46.3%	+99.3%* [+67, +141]	195	78.4%	+11.5% [−2, +22]	159
OLMo-3.1-32B-Instruct	34.7%	+109.7%* [+77, +146]	277	71.5%	+11.6% [−2, +23]	251
Llama-3.2-1B-Instruct	61.0%	+53.2%* [+33, +81]	146	91.8%	+5.6% [−1, +12]	133
Llama-3.2-3B-Instruct	60.6%	+35.0%* [+17, +53]	230	77.6%	+9.6% [−5, +21]	187
Llama-3.1-8B-Instruct	33.3%	+135.9%* [+98, +179]	255	64.9%	+44.1%* [+28, +65]	213

C WHERE THE RISK LIVES

Risk share of the per-position error. Figure C.1 gives the within-position cross-section behind the $31\% \rightarrow 49\%$ risk share of Section 3.3.

C.1 ONSET DYNAMICS

The self-readable channel is impulsive; the oracle-referenced channels persist. Re-scoring the full next-token distribution around onset on the free-run trajectories, every disagreement channel spikes together at onset, but only the oracle-referenced channels linger (Table C.1, Figure 2a): the self-readable $H(p_M)$ is impulsive, collapsing with a half-life of 0.2–0.9 tokens, while the oracle-referenced channels persist. The divergence $\mathbb{E}[\delta]$ relaxes with half-life 2.5–3.1 tokens, and the decoding risk $\sqrt{\mathrm{Var}[\delta]}$ is the longest-lived (3.4–4.6 tokens), sustaining the highest post-onset plateau of the three.

C.2 ORACLE SENSITIVITY

The onset asymmetry survives a change of oracle size. δ is divergence from a particular oracle, so a fair worry is that the persistence asymmetry is an idiosyncrasy of one reference model. It is not: re-scoring the same onset analysis against the Qwen3-14B oracle (the cross-sectional reference) instead of the Qwen3-32B oracle leaves every qualitative invariant intact (Table C.2). Under both oracles the oracle-referenced channels — divergence $\mathbb{E}[\delta]$ and decoding risk $\sqrt{\mathrm{Var}[\delta]}$ — stay elevated post-onset (the risk half-life exceeds the realized-bias half-life $\overline{\delta(y_t)}^2$ at every rung), while the self-readable $H(p_M)$ relaxes closest to baseline — the same proxy break. The two columns are scored on independently sampled free-run trajectory sets at different nucleus truncations, so the exact half-lives differ and are not expected to match; that the *asymmetry* holds across oracle size on independent data, rather than on one shared sample, is the robustness claim.

The onset asymmetry survives a change of oracle *family*. Re-scoring the same onset trajectories against a distinct-family oracle (DeepSeek-V3-671B), aligned to the Qwen3 vocabulary by prefix-routed transport, reproduces the ordering $\sqrt{\mathrm{Var}[\delta]} > \mathbb{E}[\delta] \gg H(p_M)$ (Table C.3, Figure C.2): the self-readable $H(p_M)$ stays impulsive (half-life 0.4 tokens), collapsing while the oracle-referenced channels persist. The alignment injects per-position noise that blurs the fine same-family separation — $\mathbb{E}[\delta]$ and $\sqrt{\mathrm{Var}[\delta]}$ now relax with statistically indistinguishable half-lives (1.7 vs 1.5 tokens) — so their ordering is carried by the sustained plateau instead: $\sqrt{\mathrm{Var}[\delta]}$ settles at 1.19 nats versus $\mathbb{E}[\delta]$'s 0.80. Absolute half-lives shorten against the more distant oracle, but the impulsive-versus-persistent proxy break is unchanged.

C.3 COMMITMENT REGIMES GIVE A LABEL-FREE RISK MARKER

Three commitment regimes. A per-model three-state model over the per-token triple (bias $\mathbb{E}[\delta]$, entropy $H(p_M)$, decoding risk $\sqrt{\mathrm{Var}[\delta]}$) recovers the regimes of Figure 2b: *grounded* (low bias, entropy & risk), *precarious* (low bias & entropy but high risk, a confident yet high-variance self-conditioned branch; prec below), and *diverged* (high bias, entropy & risk) — gap regimes, not correctness. The trajectory runs grounded, shifts to diverged at the verifier onset, then persists in the precarious state through the continuation — the self-conditioning signature (conditioning on its own draw) — and the gap-only fit below shows it recurs at every scale.

A gap-only fit recovers the same over-commitment signature. A three-state Markov-switching fit to the gap $\delta_t = \log p_M - \log p_O$ alone is BIC-preferred over two states for 99.9–100% of series at every scale. Its highest state is the over-commitment regime: highest mean gap and also highest variance (an over-commitment, not a low-variance "confidently-wrong" attractor), where the model over-rates the realized token relative to the oracle (Table C.4). It is reachable from the precarious state and sticky (entry and self-transition probabilities both ≈ 0.42–0.45 at every scale), and its over-rating magnitude shrinks monotonically with scale, 2.4 nats at 0.6B to 0.8 nats at 8B, matching the falling bias^2 share of Figure C.1.

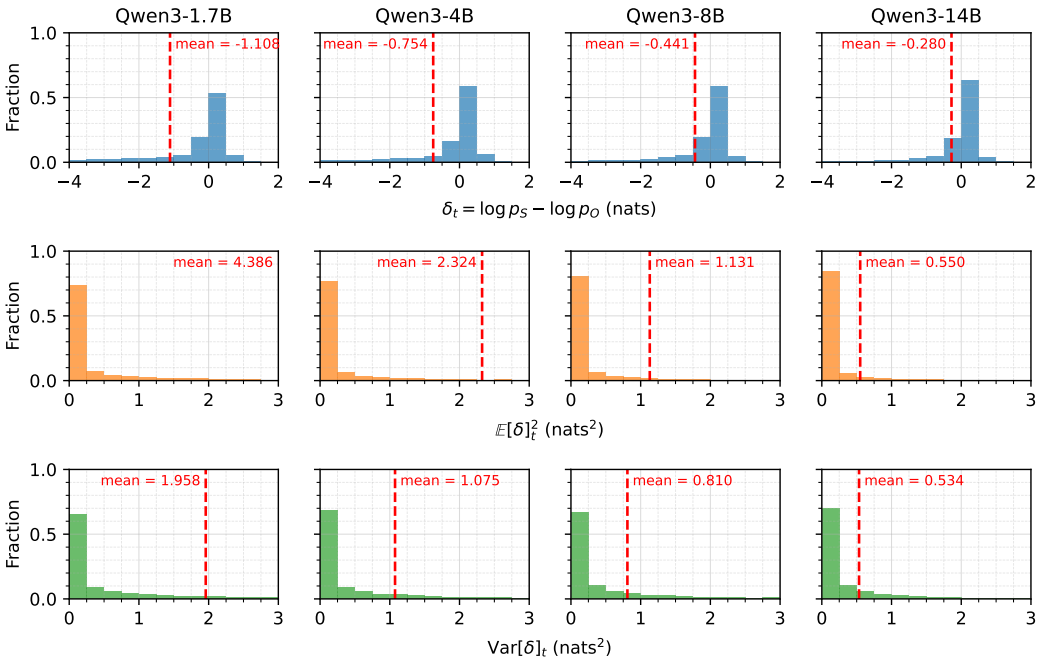


Figure C.1: **Decoding risk is a growing share of the per-position error with scale.** Teacher-forced $\delta_t = \log p_M(y_t^*) - \log p_O(y_t^*)$ (Qwen3 vs. Qwen3-32B; 14,629 tokens/model, 58 biographies; 0.6B omitted). *Top:* realized-token δ_t density, leptokurtic with scale. *Middle/bottom:* the distributional bias^2 ($\mathrm{KL}(p_M \| p_O)_t^2$) and risk ($\mathrm{Var}_{p_M}[\delta]_t$) terms (nats^2); both shrink, but the bias^2 term faster, so risk's share of $\mathbb{E}[\delta^2]_t$ rises $31\% \to 49\%$ from 1.7 to 14B. Cross-sectional view of the onset-aligned breakdown in Figure 2a.

Table C.1: **At a fabrication's onset the decoding risk is the longest-lived channel, the bias gap shorter, and the model's self-readable uncertainty the most impulsive.** The three Figure 2a channels of the disagreement variable (Equation 2) in nats — the divergence $\mathbb{E}[\delta]{=}\mathrm{KL}(p_M \| p_O)$, the decoding std $\sqrt{\mathrm{Var}[\delta]}$, and the felt uncertainty $H(p_M)$ — onset-aligned on free-running hallucinating trajectories (LongFact++; onset $t_0^\star$ = first unsupported atomic fact), the SAME frame Figure 2a plots. Models (Qwen3-0.6B–8B) scored against the Qwen3-14B oracle. Per channel the free-asymptote fit $C{+}A\,e^{-\lambda s}$ over the full post-onset window gives the pre-onset baseline ($t - t^\star \in [-5, -1]$), the onset peak, and the post-decay plateau $B{+}C$ (all nats), and the half-life $t_{1/2}{=}\ln 2/\lambda$ (tokens; $--$ where no clean exponential forms). The decoding std $\sqrt{\mathrm{Var}[\delta]}$ relaxes slowest (3.4–4.6 tokens) and sustains the highest plateau, the bias gap $\mathbb{E}[\delta]$ faster (2.5–3.1), and the self-readable $H(p_M)$ the most impulsive (0.2–0.9).

	$\mathbb{E}[\delta]$				$\sqrt{\mathrm{Var}[\delta]}$				$H(p_M)$			
Model	pre	onset	post	$t_{1/2}$	pre	onset	post	$t_{1/2}$	pre	onset	post	$t_{1/2}$
Qwen3-0.6B	0.95	2.87	1.02	3.1	1.59	2.83	1.59	3.4	1.01	2.68	1.46	0.6
Qwen3-1.7B	0.60	2.63	0.77	2.5	1.06	2.37	1.12	4.1	0.47	1.48	0.65	0.9
Qwen3-4B	0.40	1.95	0.56	2.9	0.84	1.96	0.88	4.6	0.42	1.25	0.53	0.7
Qwen3-8B	0.29	1.19	0.38	2.8	0.75	1.51	0.82	4.4	0.57	1.39	0.63	0.2

Hallucinated tokens load the over-commitment state. This backs Section 3.3: unsupported atomic-fact tokens occupy the over-commitment state $\approx 1.7\times$ as often as supported ones (0.42–0.46 vs 0.25–0.27), separating supported from unsupported content at AUROC 0.68–0.71, *stable across all four scales* (Table C.4). The state is a label-free, white-box risk marker (read off the logits with no verifier), and it agrees with the independent FActScore position fit (Appendix B) about where in a response the model is most likely to fabricate.

Table C.2: **The onset asymmetry is invariant to oracle size.** The persistence asymmetry of Table C.1 — the oracle-referenced channels (decoding risk $\mathrm{Var}[\delta]$, divergence $\mathbb{E}[\delta]$) stay elevated post-onset (> 1) while the self-readable $H(p_M)$ relaxes closest to baseline, the risk longest-lived — reproduces whether Qwen3 models (0.6–8B) are scored against the Qwen3-14B or the Qwen3-32B oracle; the decoding-risk half-life $t_{1/2}$ exceeds the realized-bias half-life $\overline{\delta(y_t)}^2$ at every rung. po/pre is the post-onset ($t - t^\star \in [1, 10]$) mean over the pre-onset baseline. *The two columns are not a controlled swap*: they come from independently sampled free-run trajectory sets at different nucleus truncations (top-k 1024 vs 2048), so the exact half-lives differ; the qualitative asymmetry does not. Agreement across oracle size on independent data is, if anything, stronger than a shared-sample swap.

	$\overline{\delta(y_t)}^2$ $t_{1/2}$		$\mathrm{Var}[\delta]$ $t_{1/2}$		$\mathrm{Var}[\delta]$ **po/pre**		$\mathbb{E}[\delta]$ **po/pre**		$H(p_M)$ **po/pre**	
Model	14B	32B	14B	32B	14B	32B	14B	32B	14B	32B
Qwen3-0.6B	1.2	0.5	3.7	0.9	1.9×	1.7×	1.7×	1.6×	1.5×	1.3×
Qwen3-1.7B	0.7	0.7	1.7	4.6	2.3×	1.8×	2.1×	1.5×	1.6×	1.2×
Qwen3-4B	0.8	0.6	4.4	1.6	2.8×	1.8×	2.4×	1.8×	1.4×	1.3×
Qwen3-8B	0.7	0.6	1.9	4.7	2.2×	2.6×	2.1×	1.9×	1.1×	1.2×

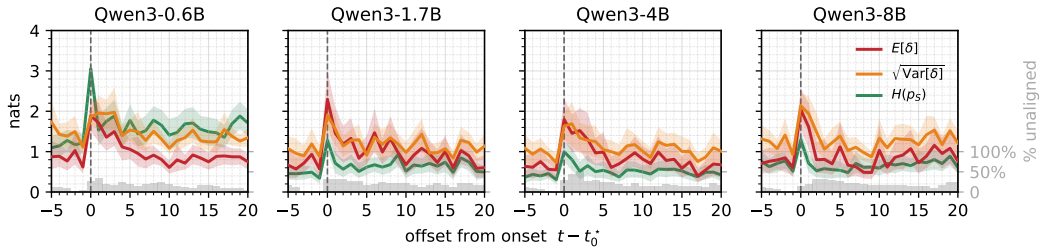


Figure C.2: **The onset dynamics reproduce against a cross-family oracle: the self-readable channel again collapses within a token while the oracle-referenced channels persist.** The same onset-aligned free-run LongFact++ trajectories as Figure 2a (Qwen3 0.6–8B, one column per rung), re-scored against a distinct-family oracle (DeepSeek-V3-671B) via vocabulary-alignment transport: the three Figure 2a nats channels — the bias gap $\mathbb{E}[\delta]{=}\mathrm{KL}(p_M \| p_O)$ (red), the decoding risk $\sqrt{\mathrm{Var}[\delta]}$ (amber), and the self-readable $H(p_M)$ (green) — as solid curves on one shared nats axis (95% CIs), the second row of Figure 2a. Per-channel half-lives and plateaus are in Table C.3. The gray rug (secondary axis) is the % of tokens not in a 1:1 exact-aligned block at each offset (a branch-point approximation of the surface-bridged channels), the cross-family alignment-reliability context; positions are pre-filtered to $\geq 80\%$ retained surface mass.

Table C.3: **The onset persistence ordering $\sqrt{\mathrm{Var}[\delta]} > \mathbb{E}[\delta] \gg H(p_M)$ survives a cross-family oracle.** The same onset-aligned LongFact++ trajectories as Table C.1 (Qwen3 0.6–8B; onset $t_0^\star$ = first unsupported atomic fact), re-scored against the cross-family DeepSeek-V3-671B oracle (a distinct model family from the Qwen3-0.6B–8B students). Per channel the free-asymptote fit $C + A\, e^{-\lambda s}$ gives the pre-onset baseline, onset peak, and post-decay plateau (nats), and the half-life $t_{1/2}$ (tokens). The self-readable $H(p_M)$ collapses fastest (half-life 0.4 tokens) while the oracle-referenced channels persist; among them $\sqrt{\mathrm{Var}[\delta]}$ sustains the higher plateau (1.19 versus 0.80 nats).

	$\mathbb{E}[\delta]$				$\sqrt{\mathrm{Var}[\delta]}$				$H(p_M)$			
Model	pre	onset	post	$t_{1/2}$	pre	onset	post	$t_{1/2}$	pre	onset	post	$t_{1/2}$
Qwen3-0.6B	0.83	1.89	0.83	2.1	1.43	1.87	1.43	2.0	1.18	3.05	1.60	0.0
Qwen3-1.7B	0.69	2.30	0.90	1.2	1.00	1.91	1.12	0.9	0.44	1.28	0.63	0.3
Qwen3-4B	0.62	1.80	0.66	2.7	0.95	1.63	1.00	2.2	0.39	1.00	0.47	1.1
Qwen3-8B	0.74	2.04	0.82	0.8	1.20	2.14	1.22	1.0	0.67	1.29	0.67	0.2

Table C.4: Unsupported tokens load the over-commitment state $\approx 1.7\times$ as often as supported ones, separating the two at AUROC 0.68–0.71 with no verifier and no labels. Three-state gap-MSAR at the BIC-selected $k{=}3$ on free-run trajectories (Qwen3 0.6–8B models, 14B oracle); $k{=}3$ is preferred over $k{=}2$ for 99.9–100% of series. The over-commitment state (highest mean gap μ_{oc}) is also the highest-variance (σ_{oc}); its over-rating magnitude shrinks with scale ($2.36 \to 0.81$ nats). AUROC is stable across all four scales.

	prefers	**over-commit state**		**occupancy**	
Model	$k{=}3$	μ_{oc}	σ_{oc}	U / S	AUROC
Qwen3-0.6B	99.9%	2.36	2.75	0.42 / 0.25	0.681
Qwen3-1.7B	100%	1.73	2.35	0.46 / 0.27	0.710
Qwen3-4B	100%	1.26	1.98	0.43 / 0.25	0.713
Qwen3-8B	100%	0.81	1.56	0.44 / 0.26	0.714

D How Risk Compounds and Hides

D.1 The Across-Claim Snowball

The across-claim snowball of Section 4.1 survives the topic-difficulty control. The relative risk $\mathrm{RR} = P(U_{j+1} \mid U_j)/P(U_{j+1} \mid S_j)$ holds in both samples. Across all responses the raw $\mathrm{RR}_{\mathrm{all}}$ climbs $\times 1.28 \rightarrow \times 2.85$ with scale; restricting to *mixed* responses (both a supported and an unsupported claim present, so the topic is at least partly known — controlling the base rate) attenuates it to $\mathrm{RR}_{\mathrm{mixed}}$ $\times 1.08 \rightarrow \times 1.71$ but does not remove it. Every 95% CI — bootstrapped over responses, the independent unit — excludes 1 on thousands of consecutive pairs per rung, so the snowball is neither a topic-gap confound nor a selection artifact of the mixed restriction (Table D.1).

Decoding ablation. The moment analyses are teacher-forced and so independent of decoding, but the free-run snowball is not. We re-run it at greedy (τ=0) and low-temperature (τ=0.5) decoding on a shared 32-prompt subset and recompute the across-claim relative risk (Table D.2). It stays above 1 at every rung and is larger at the biggest model than the smallest under all three policies, so the compounding is not an artifact of sampling temperature; the small subset adds noise at the middle rungs.

Self-grounding the first claim. Figure D.1 repeats the first-claim contrast under intervention rather than observation.

D.2 The Precarious State Bridges Adjacent Fabrications

The bridge effect reads the *precarious* state of the three-channel fit (Appendix C.3), not the gap-only over-commitment state of Table C.4: between consecutive claims, the confident precarious state fills the inter-claim bridge more when the next claim is also unsupported (Section 4.2, Figure 4). Table D.3 gives the per-rung effect at the $\leq$ 10-token adjacency window: the probability a bridge token is in the *precarious* state when the next claim is a fabrication $\Pr(\mathrm{prec} \mid \mathrm{U} \rightarrow \mathrm{U})$ vs. supported $\Pr(\mathrm{prec} \mid \mathrm{U} \rightarrow \mathrm{S})$, their absolute difference, and the ratio $\Pr(\mathrm{prec} \mid \mathrm{U} \rightarrow \mathrm{U})/\Pr(\mathrm{prec} \mid \mathrm{U} \rightarrow \mathrm{S}) - 1$, alongside the number of prev-fabrication pairs n behind each estimate. The effect is scale-emergent: strongest at the largest (14B) rung, with Qwen3-1.7B the lone negative rung. Rungs with fewer than 100 prev-fabrication pairs (the small Llama-3.2 rungs) are too sparse to estimate the ratio reliably and are excluded from Figure 4.

D.3 Detector Implementation

Semantic entropy. The detector we test is a functional of p_M alone — the object Prediction 4 concerns. The blindness test (Section 4.4, Table 2) scores each claim with the original semantic-entropy detector (Kuhn et al., 2023; Farquhar et al., 2024): the off-shelf `deberta-large-mnli` natural-language-inference model, bidirectional-entailment clustering of resamples, and the discrete cluster-proportion entropy (nats). The resamples are the K=25 same-model stochastic draws of each prompt already produced for the free-run trajectories (Appendix A). One adaptation makes it claim-level: for the claim under test, each resample's NLI stance toward the claim (premise = resample, hypothesis = claim) $\in$ {entailment, neutral, contradiction} is its semantic cluster, and the score is the discrete entropy over those clusters; the cluster count is therefore not a free hyperparameter (emergent from entailment, ≤ 3 at the claim level). Higher entropy = more resample disagreement about the claim = a stronger hallucination flag. SelfCheck-Prompt (Manakul et al., 2023), also a p_M-functional over the same draws, is implemented from its canonical leave-one-out consistency prompt.

Table D.1: **The across-claim snowball holds for both the raw and the topic-controlled sample — not a confound or a selection artifact.** Consecutive claim pairs from Qwen3 free-run trajectories (Section 4.1), shown for *all* responses and the *mixed* subset (holding both a supported and an unsupported claim, controlling the topic base rate). Per block: the consecutive-pair count n; the next claim's fabrication rate after a supported (spontaneous) vs. an unsupported claim, $\Pr(U|S)/\Pr(U|U)$; and the across-claim relative risk RR$=P(U_{j+1}|U_j)/P(U_{j+1}|S_j)$, $\pm$ half the 95% CI bootstrapped over responses (the independent unit). RR clears 1 in *both* samples at every rung, so the lift is neither a topic-gap confound nor a selection artifact of the mixed restriction; controlled, RR_{mixed} climbs $1.08 \rightarrow 1.71$ with scale.

	all responses				mixed responses			
Model	n	$\Pr(U\|S)$	$\Pr(U\|U)$	RR	n	$\Pr(U\|S)$	$\Pr(U\|U)$	RR
Qwen3-0.6B	8,064	0.74	0.95	1.28 ± 0.07	2,416	0.74	0.80	1.08 ± 0.07
Qwen3-1.7B	6,450	0.44	0.84	1.93 ± 0.12	3,877	0.48	0.68	1.40 ± 0.08
Qwen3-4B	6,820	0.33	0.75	2.25 ± 0.15	4,972	0.39	0.60	1.54 ± 0.10
Qwen3-8B	7,171	0.23	0.65	2.85 ± 0.22	5,361	0.29	0.50	1.71 ± 0.12

Table D.2: **Reliability anti-scaling survives alternative decoding.** Across-claim snowball relative risk $\text{RR} = P(U_{j+1} \mid U_j)/P(U_{j+1} \mid S_j)$ on mixed responses (Section 4.1), per Qwen3 size, under greedy (τ=0), low-temperature (τ=0.5), and the canonical τ=1.0 run, on a shared 32-prompt subset. The relative risk stays above 1 at every rung and is larger at the biggest model than the smallest under all three decoding policies (the small subset adds noise at the middle rungs), so the across-claim snowball is not an artifact of sampling temperature.

Decoding	**Q0.6B**	**Q1.7B**	**Q4B**	**Q8B**	**Q14B**
greedy (τ=0)	$\times$1.43	$\times$2.05	$\times$2.14	$\times$1.66	$\times$1.79
τ=0.5	$\times$1.22	$\times$1.67	$\times$1.85	$\times$2.09	$\times$1.97
τ=1.0 (canonical)	$\times$1.07	$\times$1.42	$\times$1.57	$\times$1.80	–

Table D.3: **Per-rung values behind the precarious-regime bridge effect (Figure 4).** Every model rung (all families, size-ordered): the number of prev-fabrication claim pairs n, the probability that an inter-claim bridge token is in the *precarious* state when the next claim is supported $\Pr(\text{prec} \mid \text{U} \rightarrow \text{S})$ vs. a fabrication $\Pr(\text{prec} \mid \text{U} \rightarrow \text{U})$, their absolute difference $\Delta = \Pr(\text{prec} \mid \text{U} \rightarrow \text{U}) - \Pr(\text{prec} \mid \text{U} \rightarrow \text{S})$, and the plotted ratio $\Pr(\text{prec} \mid \text{U} \rightarrow \text{U})/\Pr(\text{prec} \mid \text{U} \rightarrow \text{S}) - 1$ (the Figure 4 y-axis). Adjacency window ≤ 10 tokens, mixed-trajectory restricted (as Figure 4). † Daggered rungs are excluded from Figure 4: outside the 1–14 B scale window (Qwen3-0.6B) or fewer than 100 prev-fabrication pairs (Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct).

Model	n	$\Pr(\text{prec} \mid \text{U} \rightarrow \text{S})$	$\Pr(\text{prec} \mid \text{U} \rightarrow \text{U})$	Δ	$\frac{\Pr(\text{prec}\mid\text{U}\rightarrow\text{U})}{\Pr(\text{prec}\mid\text{U}\rightarrow\text{S})} - 1$
Qwen3-0.6B†	58	0.54	0.63	+0.09	+17%
Llama-3.2-1B-Instruct†	4	0.33	1.00	+0.67	+200%
Qwen3-1.7B	130	0.47	0.45	−0.02	−4%
Llama-3.2-3B-Instruct†	9	0.40	0.76	+0.35	+89%
Qwen3-4B	163	0.36	0.38	+0.02	+5%
OLMo-3-7B-Instruct	102	0.19	0.21	+0.02	+12%
Qwen3-8B	170	0.38	0.44	+0.06	+15%
Qwen3-14B	178	0.23	0.40	+0.16	+69%

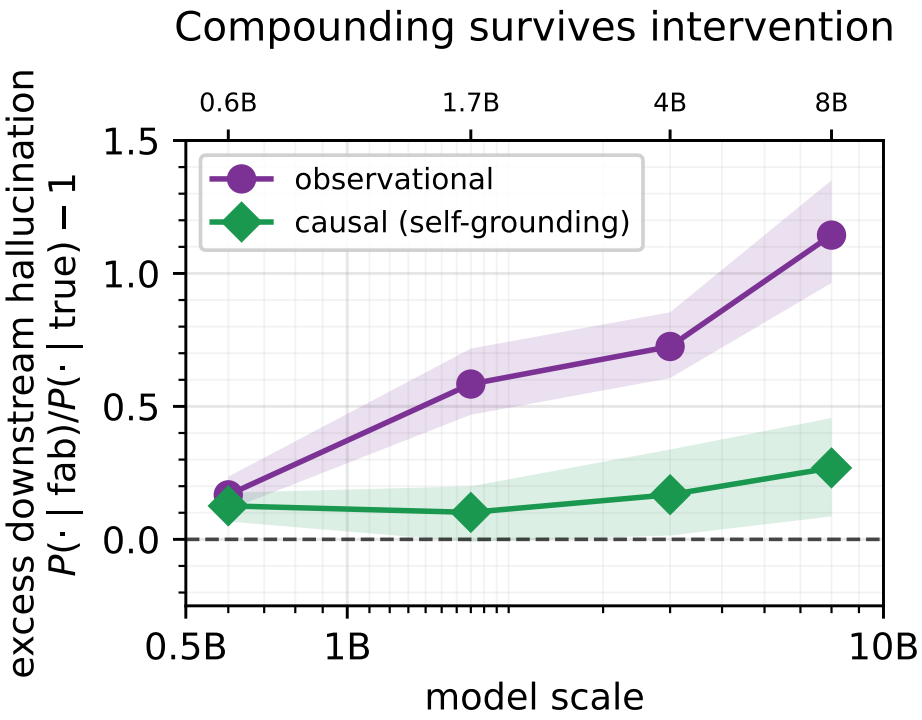


Figure D.1: **Early hallucinations causally compound downstream, though the raw association overstates the effect several-fold.** Excess downstream hallucination ($\mathrm{RR} - 1$, y) vs. scale (Qwen3 0.6–8B): the rest-of-response hallucination rate when the response's first claim is fabricated, relative to a true-first-claim baseline; above the dashed zero a fabrication *raises* the rate (shaded 95% bootstrap CIs over trajectories). Both series share one estimand and differ only in how the first claim is made true. *Observational:* responses that happened to open with a supported claim ($+0.17 \rightarrow +1.14$) — an association, with the topic-level knowledge gap as a common cause. *Causal:* the *same* trajectory's continuation after its first fabricated claim is self-grounded (rejection-sample to supported, then free-run on-policy), holding the topic fixed ($+0.13 \rightarrow +0.27$; CI excludes 0 at $3/4$). The compounding survives intervention — it follows from a sampling event, not only from missing knowledge — but the observational reading exceeds the causal one at every rung, so the association is not a substitute for the intervention.

E THE VARIANCE-CONTRACTION INTERVENTION

This section details the on-policy intervention behind Section 4.3 and reports three results the main text compresses: *where* bias and variance concentrate (the live divergence event, not the verifier label), how the contraction localizes to the inter-claim bridge, and a dose-response that pins the lever to variance. The intervention runs on the LongFact++ free-run trajectories (Appendix A) for Qwen3 (4–14B, oracle 32B), OLMo-3-7B (oracle 32B) and Llama-3.2-1B/3B (oracle 8B).

The mean-preserving contraction. At each step the co-resident oracle supplies the full next-token gap $\delta(x) = \log p_M(x) - \log p_O(x)$, its mean $\mu = \mathbb{E}_{p_M}[\delta] = \mathrm{KL}(p_M \| p_O)$ (the *bias*), and its spread $\mathrm{Var}_{p_M}[\delta]$ (the *decoding risk*); the intervention replaces p_M with the contraction q_λ of Equation 4 (main text). The penalty acts on logits via δ, so the knob is invariant to a uniform shift of $\log p_M$ and reduces to ordinary temperature sharpening only in the degenerate case $\delta \equiv$ const; otherwise it is strictly a second-moment operation. The dose ρ (λ chosen per step so $\mathrm{Var}_{q_\lambda}[\delta] = \rho\,\mathrm{Var}_{p_M}[\delta]$) is comparable across models where a raw λ is not. Holding the mean discards the *lucky-good* excursions (δ on the supported side of μ) symmetrically with the bad ones; removing only the bad ones would improve bias and amount to leaking the oracle's answer, which Equation 4 deliberately does not do.

Solving ν and λ. Both multipliers are found by one-dimensional root-finding over the model nucleus support, exploiting that each target moment is monotone in its knob. Inner solve: for a fixed penalty strength λ, $\mathbb{E}_{q_\lambda}[\delta]$ is strictly increasing in ν, so the unique ν with $\mathbb{E}_{q_\lambda}[\delta] = \mu$ is bracketed by doubling outward from a spread-scaled guess and refined by Brent's method (tolerance 10^{-10}, ≤ 200 iterations); if the constraint already holds within 10^{-2} nats the tilt is skipped (ν=0). Outer solve: $\mathrm{Var}_{q_\lambda}[\delta]$ is monotone decreasing in λ (with ν re-solved inside at each λ), so the λ meeting the dose ρ is bracketed and Brent-refined the same way. At a degenerate position — a near-collapsed support where a moment is unreachable — the solver caps at the bracket endpoint and records the realized $\mathrm{var_ratio} = \mathrm{Var}_{q_\lambda}[\delta]/\mathrm{Var}_{p_M}[\delta]$ and mean drift honestly rather than aborting the co-resident generate barrier; the manipulation check (Figure E.1) reports these realized moments. Cost is a handful of Brent iterations of an $O(|\mathrm{nucleus}|)$ kernel per fired position, and the controller fires only at the Z=2 trigger crossings (*not* every token), so the per-token overhead is negligible against the model+oracle forward pass it rides on.

Trigger and grid. Let $x_t^\star = \arg\max_x p_O(x)$ be the oracle's preferred token and $s_t = \log p_O(x_t^\star) - \log p_M(x_t^\star) \geq 0$ the amount the model under-weights it. The controller fires where s_t spikes Z standard deviations above its per-model mean ($s_t \geq \bar{s} + Z\sigma_s$, Z=2), standardized per model so one threshold transfers across scales; anchoring at $x^\star$ (rather than $\arg\max_x \delta$, dominated by the model's near-zero-probability tail) makes the signal "oracle confident about a token the model misses," which pre-selects the recoverable knows-but-hallucinates regime. Around each trigger a tophat window of W=2 tokens opens at offset K and applies q_λ at constant dose ρ. The swept grid is $K \in \{-1, 0, 1, 2\} \times \rho \in \{0.5, 0.75, 1.25, 1.5\}$, with the controller refractory-gated so that it fires at every crossing (M=∞ in the arm tags). The control is no-op (λ=0, the model's own CRN-identical free-run), giving one unambiguous paired baseline. We report the best arm over the *causal* $K \geq 0$ subgrid only: a window opening before the online trigger (K=-1) cannot be realized by any controller that detects the divergence before acting, so it is excluded even from the oracle-in-the-loop ceiling. Generate is co-resident on a single H100 (model+oracle, full vocabulary); verification is the same post-barrier, by-prompt web-search pipeline as the observational tests (Appendix A).

Bias and variance concentrate at the trigger, not the verifier onset. Per-position bias^2 and decoding risk are stored for every draw and plotted against two anchors: the verifier fabrication onset $t_0^\star$ (first unsupported-claim token) and the live oracle trigger t_{trig}. The bias^2 excursion forms a sharp spike at offset 0 in trigger coordinates but is smeared and off-zero in onset coordinates: the verifier label *lags* the model's computational divergence by several tokens. Any intervention keyed to the verifier onset therefore aims downstream of where bias and variance live, which is why relocating the anchor onto the live divergence event is what makes the contraction land. One caveat bounds this: aligning on a threshold crossing mechanically shapes both moments near offset 0, so we report the ordering of the two spikes without testing it against a shuffled-trigger null.

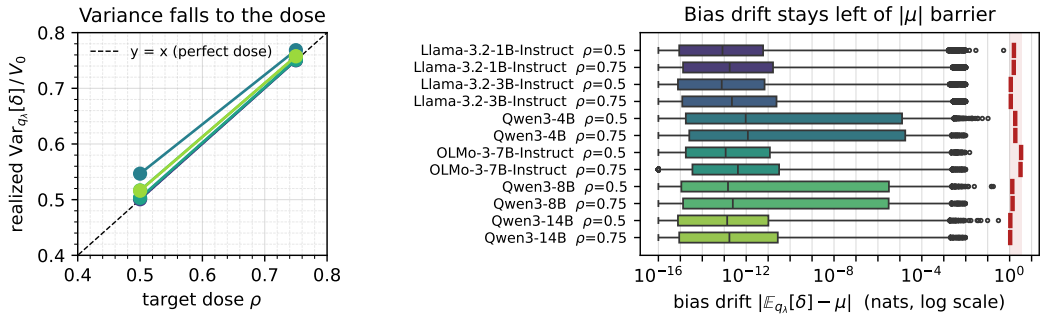


Figure E.1: **Manipulation check: the contraction moves variance and nothing else.** Left: realized per-fired-step variance fraction $\mathrm{Var}_{q_\lambda}[\delta]/V_0$ vs target dose ρ, per rung — points on the identity mean variance fell to the dose (realized 0.50–0.77 at ρ=0.5–0.75). Right: bias drift $|\mathbb{E}_{q_\lambda}[\delta] - \mu|$ (box = IQR, whiskers = 5–95%) against the $|\mu|$ barrier (red) the contraction must not cross. Median drift is $\leq 10^{-11}$ nats, $\sim$11 orders below $|\mu|$, demonstrating that the downstream effect (Table E.2) traces to decoding risk, not a mean shift or topic difficulty.

Localizing the contraction to the inter-claim bridge. The causal claim of Section 4.3 is early→late, so we relocate the same contraction onto the bridge: the inter-claim gap $[c_{k^\star}.t_1,\ c_{k^\star+1}.t_0)$ of the first consecutive-unsupported claim pair, fired at every gap token (structural — no trigger, no threshold), with the single next claim $c_{k^\star+1}$ as the outcome. The anchor is the verifier's claim labels and token spans on the control draw, which the co-resident generate cannot know online, so the run is two-pass: pass one generates and verifies the controls; pass two branches the arms from those spans at $\rho \in \{0.5, 0.75\}$ and verifies. A draw qualifies on a run of ≥ 2 consecutive unsupported claims plus ≥ 1 supported claim; intersecting the qualifying prompts across a family's students yields per-family bridge sets — 25 of the 88 pooled prompts for Qwen (16–30 control draws per student, fewer at scale as bigger models fabricate less), and comparably sized sets for OLMo and Llama (29–40 control draws) — with 15–31 verifiable arm $c_{k^\star+1}$ per cell. The verifier force-searches every arm $c_{k^\star+1}$ and skips the arm's other claims — the prefix $c_{\leq k^\star}$ is measured on the control, and claims past $c_{k^\star+1}$ are outside the outcome — so no outcome claim is dropped unassessed (Table E.1). Everything else — kernel, common random numbers, calibration, verification, and the trajectory-clustered paired bootstrap — is the whole-response protocol above, verbatim.

Dose-response: contraction helps, expansion does not. At $M{=}\infty$, $K{=}0$, the paired Δ is positive for every contraction dose ($\rho{<}1$) at all six rungs. On the four Qwen3 and OLMo rungs it is near-zero or negative for every expansion dose ($\rho{>}1$); this contraction-vs-expansion gap (deeper variance suppression helps, variance inflation does not) is the robust signal that the lever is variance. The ρ=1.25 and ρ=1.5 cells are identical because the kernel clamps the expansion side; on the two Llama-3.2 rungs that clamp lands on a contraction, so their nominal expansion cells also reduce the rate ($\Delta \approx 0.21/0.26$) and provide no clean expansion control. The ordering *within* the contraction side (ρ=0.5 vs 0.75) is model-specific. This is the dose-response image of Equation 4 acting on the variance term of Equation 2 at fixed KL.

Absolute effects. Table E.2 carries the absolute rates behind the relative reduction of Figure 1b: the best-arm paired Δ is 0.13–0.33 with a 95% bootstrap CI excluding zero at all six model×family rungs, and the winning arm always sits on the contraction side ($\rho{\leq}0.75$). The manipulation check (Figure E.1) confirms the realized variance falls to the dose while the per-fired-step bias drift stays $\sim$11 orders of magnitude below $|\mu|$, so the reduction is attributable to the decoding-risk channel and not to a residual mean shift.

No length, claim-count, or informativeness confound. The contraction changes *which* claims a response supports, not how many it makes or how long it runs. Table E.3 holds each rung's best arm against control on four descriptive statistics: length is pinned by the generation cap (within 0.5%), fixing the rate's denominator; the atomic-claim count — the informativeness signal at fixed length — moves non-directionally within -9.7% to $+4.2\%$; and self-BLEU (stylistic diversity, normalized

Table E.1: **Contracting the inter-claim bridge prevents the second fabrication (absolute effects).** Per student and dose ρ: the verifiable (n) count and web-verified unsupported rate of the second fabrication's slot $c_{k^\star+1}$, the within-trajectory paired Δ (control $-$ arm) with its trajectory-clustered 95% bootstrap CI, and the arm's Supported/Not-Supported/Insufficient-Information label counts. The control rate is 1.0 by construction — the bridge is the inter-claim gap of a *consecutive-unsupported pair*, so the control's second claim is unsupported — and Δ is therefore the fraction of second fabrications the contraction prevents. A bullet (•) marks a CI excluding zero: $\Delta = 0.21$–0.60 at all fourteen student×dose cells ($n = 15$–31). Every arm $c_{k^\star+1}$ is force-searched by the verifier, so no claim is dropped unassessed (Appendix E). As in Table E.2, the co-resident oracle makes this the *ceiling* of an online inter-claim controller, not a deployable decoding rule.

Model	ρ	n	rate	Δ [95% CI]	S/NS/II	sig.
Qwen3-1.7B	0.50	29	0.793	0.207 [0.069, 0.379]	6/23/1	•
Qwen3-1.7B	0.75	29	0.724	0.276 [0.138, 0.448]	8/21/1	•
Qwen3-4B	0.50	26	0.500	0.500 [0.308, 0.692]	13/13/3	•
Qwen3-4B	0.75	27	0.481	0.519 [0.333, 0.704]	14/13/2	•
Qwen3-8B	0.50	18	0.667	0.333 [0.111, 0.556]	6/12/1	•
Qwen3-8B	0.75	18	0.611	0.389 [0.167, 0.611]	7/11/1	•
Qwen3-14B	0.50	15	0.533	0.467 [0.200, 0.733]	7/8/1	•
Qwen3-14B	0.75	15	0.400	0.600 [0.333, 0.800]	9/6/1	•
Olmo-3-7B	0.50	22	0.636	0.364 [0.182, 0.545]	8/14/4	•
Olmo-3-7B	0.75	21	0.524	0.476 [0.286, 0.668]	10/11/6	•
Llama-3.2-1B	0.50	31	0.742	0.258 [0.129, 0.419]	8/23/4	•
Llama-3.2-1B	0.75	28	0.750	0.250 [0.107, 0.393]	7/21/9	•
Llama-3.2-3B	0.50	27	0.630	0.370 [0.185, 0.556]	10/17/12	•
Llama-3.2-3B	0.75	31	0.613	0.387 [0.226, 0.548]	12/19/9	•

Table E.2: **Contracting decoding risk causally lowers downstream hallucination (absolute effects).** Per rung: the no-op control and best-contraction-arm web-verified unsupported-claim rates over the rest of the response, the within-trajectory paired Δ (control $-$ best) with its trajectory-clustered 95% bootstrap CI, and the winning arm's onset offset K and dose ρ. A bullet (•) marks a CI that excludes zero: $\Delta = 0.13$–0.33 at all six model×family rungs. The best arm is selected post hoc per rung over the causal ($K \geq 0$) grid, so Δ is the oracle-in-the-loop *ceiling* of an online risk controller, not a deployable decoding rule. The relative reduction is the page-1 teaser (Figure 1a).

Model	ctrl.	best	Δ [95% CI]	K	ρ	sig.
Qwen3-4B	0.374	0.219	0.158 [0.020, 0.313]	+1	0.5	•
Qwen3-8B	0.346	0.128	0.221 [0.131, 0.312]	+2	0.5	•
Qwen3-14B	0.219	0.088	0.131 [0.054, 0.202]	0	0.5	•
OLMo-3-7B-Instruct	0.606	0.394	0.211 [0.041, 0.391]	+1	0.5	•
Llama-3.2-1B	0.544	0.216	0.247 [0.125, 0.360]	+1	0.5	•
Llama-3.2-3B	0.457	0.119	0.332 [0.261, 0.399]	+1	0.5	•

to $[0, 1]$) shifts a mean $+1.0\%$, up on three rungs and down on three. Only the firing rate moves decisively (zero for control). A sub-10% change in claim count at fixed length cannot produce a 35–74% drop in the unsupported *fraction*: the effect is orthogonal to, and far larger than, any descriptive shift.

Table E.3: **The hallucination reduction is not an artifact of shorter or sparser responses.** Per rung, the no-op control and the best-contraction arm (the same $K \geq 0$, $M{=}\infty$ selection as Table E.2) on three descriptive statistics of the free-run response: length (*tokens*, fixed by the generation cap), emitted atomic-claim count (*claims* — at fixed length, also the informativeness signal), and stylistic diversity (*self-BLEU*; higher = less diverse). The best arm shifts the emitted-claim count by at most 10% and self-BLEU by at most 0.021 relative to control — neither distinguishable from no intervention — while the controller *fires* on 16–48 tokens per draw (*fires*, 0 for control). A model that fabricated less by emitting fewer or shorter claims would register here; none does, so the 35–74% reduction (Table E.2) is a smaller unsupported *fraction* at fixed, equally informative output, attributable to the contracted decoding-risk channel.

Model	arm	n	tokens	claims	self-BLEU	fires
Llama-3.2-1B	control	86	254	6.69	0.103	0.0
	$K{=}+1,\ \rho{=}0.5$	86	253	6.17	0.098	47.9
Llama-3.2-3B	control	84	256	8.30	0.138	0.0
	$K{=}+1,\ \rho{=}0.5$	84	255	7.91	0.139	29.4
OLMo-3-7B	control	22	256	7.68	0.097	0.0
	$K{=}+1,\ \rho{=}0.5$	22	256	7.32	0.091	29.7
Qwen3-4B	control	21	256	8.52	0.118	0.0
	$K{=}+1,\ \rho{=}0.5$	21	256	8.19	0.108	19.9
Qwen3-8B	control	23	256	8.48	0.144	0.0
	$K{=}+2,\ \rho{=}0.5$	23	256	7.65	0.151	16.0
Qwen3-14B	control	20	256	8.25	0.106	0.0
	$K{=}0,\ \rho{=}0.5$	20	256	8.60	0.127	15.5