From Rational Correction to Hallucinated Repair: The Nishimori Manifold of Noisy-Channel Sentence Processing

Anonymous ACL submission

Abstract

Noisy-channel theories explain why comprehenders sometimes repair corrupted or underinformative utterances, but they do not identify the boundary between rational correction and hallucinated repair. We cast sentence processing as a teacher–student noisy-channel inference problem and introduce a Bayes-matched *Nishimori manifold* for sentence repair. On this manifold, we prove a sentence-level Nishimori identity: conditional on the observed utterance and discourse context, the planted intended sentence is exchangeable with an additional posterior replica. As corollaries, the replica–truth overlap gap vanishes exactly and posterior variance equals Bayes risk. We then introduce a recursive repair grammar whose posterior maps exactly to a random-field spin system on a derivation tree. For this family, we derive an exact finite-depth susceptibility $\chi_d = \sum_{\ell=0}^{d}(b \tanh \hat{J})^{\ell}$ and prove the threshold law $\Delta_c(d) = |u_0|/\chi_d + O(|u_0|^3/\chi_d^3)$. Exact synthetic experiments verify the matched-manifold identities and the threshold law, while a symbolic benchmark shows that the same mechanism remains predictive on human-readable omission-sensitive inputs. Together, these results identify Bayes matching and recursive susceptibility as organizing principles for hallucinated repair in noisy-channel language processing.

1 Introduction

Noisy-channel theories of language comprehension model interpretation as Bayesian inference over an intended utterance observed through a potentially corrupting channel. In this view, anomaly correction and context-sensitive reinterpretation arise because the comprehender trades off the literal surface form against prior expectations about what a speaker likely intended and how that intention may have been distorted in transmission (Levy, 2008, 2011; Gibson et al., 2013). Later work shows that these inferences are shaped by noisy context, adaptation to the local noise ecology, and algorithmic constraints on approximate posterior computation (Futrell and Levy, 2017; Ryskin et al., 2018; Chen et al., 2023; Clark et al., 2025). But one question remains unresolved: *where is the boundary between rational correction and hallucinated repair*? Existing noisy-channel models explain why repair can occur, but they do not identify a distinguished calibration regime or a sharp law for when small mismatch between the true corruption process and the comprehender's assumed one triggers systematic false reinterpretation.

We answer this question by importing a specific idea from Bayes-matched statistical physics. In planted inference problems, Nishimori's theory identifies a matched manifold on which planted states and posterior replicas are exchangeable and exact identities follow (Nishimori, 2001). We show that noisy-channel sentence processing admits an analogous structure. Concretely, we cast comprehension as a teacher–student problem: a teacher generates a discourse context, an intended latent sentence state, and an observed surface form under a corruption channel; a student then performs Bayesian inference with a possibly mismatched prior and channel. This yields a new theoretical question for NLP: what exact identities hold on the Bayes-matched manifold, and how does recursive sentence structure amplify off-manifold mismatch into false repair?

Our first result is a sentence-level Nishimori identity. On the Bayes-matched manifold, conditional on the observed utterance and context, the planted intended sentence is exchangeable with an additional posterior replica. Two consequences follow immediately: the *Nishimori gap* between replica–replica and truth–replica overlap vanishes exactly, and posterior variance equals Bayes risk. Thus Bayes matching is not merely a convenient parameter setting; it is the exact calibration regime of sentence repair.

Our second result characterizes how that calibration fails. We introduce a recursive repair grammar in which latent repair decisions interact across clause levels and show that its posterior maps exactly to a random-field spin system on a derivation tree. From the exact cavity recursion, we derive a finite-depth susceptibility

$$\chi_d = \sum_{\ell=0}^{d} (b \tanh(\hat{J}))^{\ell}, \tag{1.1}$$

and prove the threshold law

$$\Delta_c(d) = \frac{|u_0|}{\chi_d} + O\left(\frac{|u_0|^3}{\chi_d^3}\right). \tag{1.2}$$

Recursive syntax therefore acts as a *susceptibility amplifier*: in the subcritical regime hallucinated repair requires $O(1)$ mismatch, on the critical line the threshold scales as $1/d$, and in the supercritical regime exponentially small mismatch can already trigger a repaired interpretation that was never intended.

Our third result is empirical. Exact synthetic experiments verify the matched-manifold identities and the threshold law under exact posterior inference, and a symbolic benchmark shows that the same mechanism remains predictive on sentence-like omission-sensitive inputs. Taken together, the paper introduces a Bayes-matched manifold for noisy-channel sentence processing, derives exact calibration and threshold laws, and identifies recursive susceptibility as the quantity that organizes hallucinated repair.

2 Teacher–Student Noisy-Channel Framework

Noisy-channel theories of sentence processing model comprehension as Bayesian inference over an intended utterance observed through a potentially corrupting channel (Levy, 2008, 2011; Gibson et al., 2013; Futrell and Levy, 2017). More recent work shows that inferred repairs depend not only on local surface evidence, but also on discourse context, the assumed noise ecology, and algorithmic constraints on inference (Ryskin et al., 2018; Chen et al., 2023; Clark et al., 2025). Our goal is narrower and more structural: we isolate the effect of *model mismatch* itself, independently of approximate inference. To do so, we cast sentence processing as a teacher–student inference problem and define its Bayes-matched manifold in the sense of Nishimori (Nishimori, 2001).

Let $\mathcal{Z}$ denote a finite or countable latent sentence space. A state $z \in \mathcal{Z}$ may encode the intended string together with any hidden syntactic, semantic, or repair variables needed by the model; at this stage we keep z abstract and instantiate it in Section 4. The *teacher* draws discourse context C, intended latent state $Z^\star$, and observed surface form X according to

$$\begin{aligned} C &\sim \mu, \\ Z^\star &\sim P_\theta(\cdot \mid C), \\ X &\sim Q_\eta(\cdot \mid Z^\star, C). \end{aligned} \tag{2.1}$$

The *student* or comprehender uses a possibly mismatched prior–channel pair $(P_{\hat{\theta}}, Q_{\hat{\eta}})$, inducing the posterior

$$\begin{aligned} \Pi_{\hat{\theta},\hat{\eta}}(z \mid x, c) &= \frac{P_{\hat{\theta}}(z \mid c) Q_{\hat{\eta}}(x \mid z, c)}{\mathsf{Z}_{\hat{\theta},\hat{\eta}}(x, c)}, \\ \mathsf{Z}_{\hat{\theta},\hat{\eta}}(x, c) &:= \sum_{z' \in \mathcal{Z}} P_{\hat{\theta}}(z' \mid c) Q_{\hat{\eta}}(x \mid z', c). \end{aligned} \tag{2.2}$$

Equation (2.2) is deliberately exact: throughout Sections 3–5 we study the posterior geometry induced by mismatch before introducing any algorithmic approximation.

Assumption 2.1 (Well-posed latent sentence model)**.** The latent space $\mathcal{Z}$ is finite or countable, and for every (x, c) in the support of (X, C), the normalizer $\mathsf{Z}_{\hat{\theta},\hat{\eta}}(x, c)$ is finite and strictly positive.

Definition 2.2 (Nishimori manifold and posterior replicas)**.** The *Nishimori manifold* is the Bayes-matched subset

$$\mathcal{N} := \{(\hat{\theta}, \hat{\eta}) : \hat{\theta} = \theta,\ \hat{\eta} = \eta\}. \tag{2.3}$$

Conditional on $(X, C) = (x, c)$, posterior replicas $Z^{(1)}, Z^{(2)}, \ldots$ are i.i.d. draws from $\Pi_{\hat{\theta},\hat{\eta}}(\cdot \mid x, c)$.

The matched manifold $\mathcal{N}$ is the sentence-processing analogue of Bayes-optimal inference in planted statistical models (Nishimori, 2001). It is the unique regime in which the student prior and corruption channel coincide with the teacher that generated the data. Our main theoretical question is what changes, quantitatively and qualitatively, when inference moves off $\mathcal{N}$.

To state the later results in a model-independent way, let $R : \mathcal{Z} \times \mathcal{Z} \to [-1, 1]$ be any bounded measurable overlap observable; in the recursive repair model of Section 4, R will specialize to normalized overlap of latent repair assignments.

We define the associated *Nishimori gap*

$$\Gamma_R := \mathbb{E}\Big[R(Z^{(1)}, Z^{(2)}) - R(Z^\star, Z^{(1)})\Big], \quad (2.4)$$

where the outer expectation is taken over the joint law induced by (2.1), (2.2), and the posterior replicas. When $\Gamma_R > 0$, the posterior is more self-consistent than truth-consistent: two independent repairs sampled from the student posterior agree with each other more than they agree with the planted intended state. Section 3 proves that Γ_R vanishes identically on $\mathcal{N}$, making it a natural order parameter for hallucinated repair under mismatch.

Finally, for any bounded observable $\phi : \mathcal{Z} \to \mathbb{R}$, we write

$$\langle \phi(Z) \rangle := \sum_{z \in \mathcal{Z}} \phi(z) \Pi_{\hat{\theta},\hat{\eta}}(z \mid X, C)$$

for posterior expectation under the student model. For binary diagnostics $\psi : \mathcal{Z} \to \{-1, +1\}$, we use the Bayes decision rule

$$\widehat{\psi}(X, C) = \operatorname{sign}\big(\langle \psi(Z) \rangle\big), \quad (2.5)$$

with ties broken in favor of -1. In Sections 4–5, ψ will be the root repair variable of a recursive grammar, and the central phase-transition question will be when $\widehat{\psi}$ flips from literal interpretation to repaired interpretation as teacher–student mismatch increases.

3 The Nishimori Manifold and Exact Identities

The noisy-channel models of Levy (2008, 2011) define sentence comprehension as Bayesian inference under uncertainty, while later work studies how that inference changes under contextual modulation, adaptive noise expectations, or algorithmic resource constraints (Futrell and Levy, 2017; Ryskin et al., 2018; Chen et al., 2023; Clark et al., 2025). Our starting point is more structural. We first isolate the *Bayes-matched* regime, where the student prior and corruption channel coincide with the teacher that generated the observation, and characterize the exact identities that hold there. These identities are the sentence-processing analogue of the classical Nishimori identities in Bayesian statistical physics (Nishimori, 2001).

Throughout this section we work on the matched manifold $\mathcal{N}$, so that $(\hat{\theta}, \hat{\eta}) = (\theta, \eta)$. In that regime, the student posterior coincides with the conditional law of the planted intended state:

$$\begin{aligned} \mathsf{Z}_{\theta,\eta}(x, c) &:= \sum_{z' \in \mathcal{Z}} P_\theta(z' \mid c)\, Q_\eta(x \mid z', c), \\ \Pi_{\theta,\eta}(z \mid x, c) &= \mathbb{P}(Z^\star = z \mid X = x, C = c) \\ &= \frac{P_\theta(z \mid c)\, Q_\eta(x \mid z, c)}{\mathsf{Z}_{\theta,\eta}(x, c)}. \end{aligned} \quad (3.1)$$

Equation (3.1) is the entire source of the identities below: conditional on (X, C), the planted state $Z^\star$ and any posterior replica have the same law. The resulting exchangeability is exact and does not rely on asymptotics, Gaussian approximations, or replica methods.

Theorem 3.1 (Sentence Nishimori identity)**.** *Assume Assumption 2.1. Let $k \geq 1$, and let $Z^{(1)}, \ldots, Z^{(k+1)}$ be posterior replicas as in Definition 2.2. Then, for any bounded measurable observable*

$$f : \mathcal{Z}^{k+1} \times \mathcal{X} \times \mathcal{C} \to \mathbb{R},$$

the following identity holds on the matched manifold $\mathcal{N}$:

$$\begin{aligned} F_1 &:= f\Big(Z^\star, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big), \\ F_2 &:= f\Big(Z^{(k+1)}, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big), \\ \mathbb{E}[F_1] &= \mathbb{E}[F_2]. \end{aligned} \quad (3.2)$$

Proof sketch. Condition on (X, C). By (3.1), the planted state $Z^\star$ has conditional law $\Pi_{\theta,\eta}(\cdot \mid X, C)$. By construction, the replicas $Z^{(1)}, \ldots, Z^{(k+1)}$ are i.i.d. draws from exactly the same posterior, conditionally on (X, C). Hence the tuple $(Z^\star, Z^{(1)}, \ldots, Z^{(k)})$ is conditionally exchangeable with $(Z^{(k+1)}, Z^{(1)}, \ldots, Z^{(k)})$. Taking expectations over (X, C) yields (3.2). The complete proof is given in Appendix A.

Theorem 3.1 says that, on $\mathcal{N}$, the planted intended sentence is statistically indistinguishable from an additional posterior draw once the observed utterance and discourse context are fixed. For sentence processing, this is the exact calibration regime: posterior self-consistency and truth consistency coincide.

Corollary 3.2 (Zero Nishimori gap)**.** *Let $R : \mathcal{Z} \times \mathcal{Z} \to [-1, 1]$ be any bounded measurable overlap observable. Then, on $\mathcal{N}$,*

$$\mathbb{E}\Big[R(Z^\star, Z^{(1)})\Big] = \mathbb{E}\Big[R(Z^{(1)}, Z^{(2)})\Big]. \quad (3.3)$$

Equivalently, the Nishimori gap defined in (2.4) *vanishes:*

$$\Gamma_R = 0. \tag{3.4}$$

Proof sketch. Apply Theorem 3.1 with $k = 1$ and $f(z_0, z_1, x, c) = R(z_0, z_1)$. Since posterior replicas are conditionally i.i.d., $(Z^{(1)}, Z^{(2)})$ is exchangeable, yielding (3.3). Appendix A gives the full derivation.

Corollary 3.2 identifies Γ_R as a natural order parameter for hallucinated repair. If $\Gamma_R > 0$, the student posterior is more self-consistent than truth-consistent: two independent repairs sampled from the posterior agree with each other more than either agrees with the planted intended state. The corollary shows that exact Bayes-matched inference cannot generate such a gap; any nonzero Γ_R must arise from mismatch, approximation, or both. In this paper we isolate mismatch by keeping inference exact in the controlled model and in the synthetic validations.

Corollary 3.3 (Bayes risk equals posterior variance)**.** *Let $\phi : \mathcal{Z} \to \mathbb{R}$ be square-integrable. Then, on $\mathcal{N}$,*

$$V_\phi := \langle \phi(Z)^2 \rangle - \langle \phi(Z) \rangle^2 ,$$
$$\mathbb{E}\Big[\big(\phi(Z^\star) - \langle \phi(Z) \rangle\big)^2\Big] = \mathbb{E}[V_\phi]. \tag{3.5}$$

Proof sketch. Condition on (X, C). Under (3.1), the planted state $Z^\star$ has the posterior law, so the conditional mean-squared error of the Bayes estimator $\langle \phi(Z) \rangle$ is exactly the posterior variance of $\phi(Z)$. Averaging over (X, C) gives (3.5). The full proof is deferred to Appendix A.

Corollary 3.3 is the uncertainty-calibration counterpart to Corollary 3.2. On the Nishimori manifold, posterior uncertainty is neither heuristic nor merely introspective: it is exactly equal to Bayes risk. Taken together, Theorem 3.1 and Corollaries 3.2–3.3 identify $\mathcal{N}$ as the unique regime in which sentence repair is perfectly calibrated to the true corruption ecology. The rest of the paper studies what fails away from that manifold. In Sections 4 and 5, we show that once recursive sentence structure is introduced, even a small teacher–student mismatch can be amplified into a macroscopic false-repair effect.

4 Recursive Repair Grammar

The identities of Section 3 are completely general, but they do not yet explain how a small teacher–student mismatch can be amplified into a large false-repair effect. To make that question analytically sharp, we introduce a controlled recursive family in which repair hypotheses interact across clause levels. Appendix B gives an explicit finite weighted grammar construction over strings; here we work with its equivalent latent tree representation, which is exact rather than heuristic.

Fix branching factor $b \geq 1$ and depth $d \geq 0$, and let $T_{b,d}$ be the rooted ordered b-ary tree of depth d with vertex set $V_{b,d}$ and edge set $E_{b,d}$. Each node $v \in V_{b,d}$ carries a binary latent variable

$$s_v \in \{-1, +1\},$$

where $s_v = -1$ denotes the literal hypothesis and $s_v = +1$ denotes the repaired hypothesis. Linguistically, the nodes represent recursively embedded repair-sensitive sites, such as clause boundaries or low-information structural markers whose omission or insertion can trigger reinterpretation. The teacher prior over repair assignments is

$$P_\theta(s \mid c) \propto \exp\left(J \sum_{(u,v) \in E_{b,d}} s_u s_v + h(c) \sum_{v \in V_{b,d}} s_v \right), \tag{4.1}$$

where J couples repair decisions across parent–child sites and $h(c)$ is a context-dependent field. Positive J is the relevant regime for recursive consistency: if a clause is locally biased toward repair, its descendants become easier to repair as well.

Conditional on s, each node emits a local cue $y_v \in \mathcal{Y}$ independently through a structure-sensitive channel

$$Q_\eta(y \mid s, c) = \prod_{v \in V_{b,d}} q_\eta(y_v \mid s_v, c), \tag{4.2}$$

where $y = (y_v)_{v \in V_{b,d}}$. The full observed sentence x is obtained by a deterministic linearization of the cue field y along the recursive derivation tree. Appendix B proves that this linearization is injective, so conditioning on x is equivalent to conditioning on y. The family is therefore a genuine weighted-grammar noisy-channel model in the sense of Levy (2008, 2011), but with explicit recursive couplings between repair variables.

The key structural fact is that the student posterior over repair assignments is exactly a random-field spin system.

Theorem 4.1 (Exact posterior mapping)**.** *Assume $\hat{q}_{\hat{\eta}}(y \mid \sigma, c) > 0$ for all $y \in \mathcal{Y}$, $\sigma \in \{-1, +1\}$, and*

$c \in \mathcal{C}$. For the recursive repair grammar $\mathcal{G}_{b,d}$, the student posterior over latent repair assignments satisfies

$$\Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) \propto \exp\left(\begin{array}{l}\hat{J} \sum_{(u,v)\in E_{b,d}} s_u s_v \\ \quad + \sum_{v\in V_{b,d}} u_{\hat{\eta}}(y_v, c)\, s_v\end{array}\right). \tag{4.3}$$

where the effective local field is

$$u_{\hat{\eta}}(y_v, c) = \hat{h}(c) + \frac{1}{2}\log\frac{\hat{q}_{\hat{\eta}}(y_v \mid +1, c)}{\hat{q}_{\hat{\eta}}(y_v \mid -1, c)}. \tag{4.4}$$

Equivalently, conditioned on the observed sentence x, the posterior is exactly the same Gibbs distribution after identifying x with its unique cue field y.

Proof sketch. Appendix B constructs an explicit finite weighted grammar whose complete derivation weight factorizes into a tree prior over s and a product of local channel terms. Bayes' rule then gives

$$\Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) \propto \exp\left(\hat{J}\sum_{(u,v)} s_u s_v + \hat{h}(c)\sum_v s_v\right) \times \prod_v \hat{q}_{\hat{\eta}}(y_v \mid s_v, c).$$

Each local likelihood term is exactly affine in s_v after taking logarithms, which yields (4.3)–(4.4). No approximation is involved.

Theorem 4.1 is the point at which the paper becomes more than a generic noisy-channel model. The recursive grammar induces a posterior geometry with nontrivial couplings across repair sites, while the observed cues enter only through local log-odds fields. This gives a precise decomposition of sentence repair into *recursive consistency* and *local evidence*, making the subsequent susceptibility analysis exact rather than metaphorical.

For the critical analysis in Section 5, we specialize to the homogeneous-field regime in which context is fixed and all observed cues are of the same type, so that

$$u_{\hat{\eta}}(y_v, c) \equiv u \qquad \text{for all } v \in V_{b,d}. \tag{4.5}$$

In that case the posterior reduces to

$$\Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) \propto \exp\left(\begin{array}{l}\hat{J} \sum_{(u,v)\in E_{b,d}} s_u s_v \\ \quad + u \sum_{v\in V_{b,d}} s_v\end{array}\right). \tag{4.6}$$

with $\rho = \tanh(\hat{J})$ controlling the strength of recursive coupling. The Bayes decision at the root is

$$\widehat{\psi}(y, c) = \operatorname{sign}\big(\langle s_\varnothing \rangle\big),$$

where $s_\varnothing$ is the repair variable at the root of $T_{b,d}$.

Finally, fix a clean ambiguous cue type y° and context c°, and define the teacher-consistent literal bias

$$u_0 := h(c^\circ) + \frac{1}{2}\log\frac{q_\eta(y^\circ \mid +1, c^\circ)}{q_\eta(y^\circ \mid -1, c^\circ)}. \tag{4.7}$$

We will work in the regime $u_0 < 0$, meaning that under the true teacher model the clean ambiguous input still favors the literal interpretation. Off the Nishimori manifold, the student sees an effective field

$$u = u_0 + \Delta, \tag{4.8}$$

where the mismatch shift is

$$\begin{aligned}\Delta h^\circ &:= \hat{h}(c^\circ) - h(c^\circ),\\ \Delta q^\circ &:= \log\frac{\hat{q}_{\hat{\eta}}(y^\circ \mid +1, c^\circ)}{\hat{q}_{\hat{\eta}}(y^\circ \mid -1, c^\circ)} + \log\frac{q_\eta(y^\circ \mid -1, c^\circ)}{q_\eta(y^\circ \mid +1, c^\circ)},\\ \Delta &= \Delta h^\circ + \frac{1}{2}\Delta q^\circ.\end{aligned} \tag{4.9}$$

Section 5 shows that recursive structure transforms this seemingly local mismatch Δ into a depth-dependent global instability: once the susceptibility is large enough, an $O(\chi_d^{-1})$ perturbation suffices to flip the Bayes estimator from literal interpretation to repaired interpretation.

5 Critical Susceptibility and Hallucinated Repair

The central scientific question left open by noisy-channel models is not whether repair can occur, but how much mismatch is required before rational correction turns into self-consistent false repair (Levy, 2008, 2011; Clark et al., 2025). In the recursive repair grammar of Section 4, this question becomes analytically exact. Appendix C shows that the homogeneous depth-t root magnetization

$$m_t(u) := \langle s_\varnothing \rangle$$

obeys the exact recursion

$$m_0(u) = \tanh(u), \qquad m_{t+1}(u) = F_u(m_t(u)), \tag{5.1}$$

with

$$\begin{aligned}\rho &:= \tanh(\hat{J}),\\ A_u(m) &:= u + b\,\operatorname{artanh}\big(\rho\, m\big),\\ F_u(m) &:= \tanh\big(A_u(m)\big).\end{aligned} \tag{5.2}$$

Equation (5.2) cleanly separates *local evidence* u from *recursive amplification* through the coupling ρ. The remainder of this section extracts from that recursion an exact susceptibility law and an exact threshold law for hallucinated repair.

Proposition 5.1 (Monotonicity)**.** *Assume the ferromagnetic regime $\hat{J} \geq 0$, equivalently $\rho \in [0, 1)$.*

1. *For every depth $t \geq 0$, the homogeneous response $u \mapsto m_t(u)$ is odd and strictly increasing.*

2. *Fix a root-local literal bias $u_0 < 0$. Define the mixed-field root response*

$$\begin{aligned} A_{d-1}(\Delta) &:= \operatorname{artanh}\big(\rho\, m_{d-1}(\Delta)\big), \\ B_d(u_0, \Delta) &:= u_0 + \Delta + b A_{d-1}(\Delta), \\ M_d(u_0, \Delta) &:= \begin{cases} \tanh(u_0 + \Delta), & d = 0, \\ \tanh\big(B_d(u_0, \Delta)\big), & d \geq 1. \end{cases} \end{aligned} \tag{5.3}$$

Then $M_d(u_0, \Delta)$ is strictly increasing in each argument u_0 and Δ.

Proof sketch. Appendix E proves both claims directly from the exact cavity equations. In the ferromagnetic regime, $\operatorname{artanh}(\rho m)$ and $\tanh(\cdot)$ are strictly increasing, and the tree factorization prevents any sign frustration. Consequently, both the homogeneous and mixed-field responses are monotone. The oddness of m_t follows from the symmetry of the binary repair encoding under simultaneous sign flip of field and latent spins.

Proposition 5.1 has an immediate conceptual consequence: once $u_0 < 0$ is fixed, there is at most one mismatch level at which the Bayes decision at the root changes sign. Hallucinated repair is therefore governed by a sharp scalar threshold rather than by diffuse qualitative behavior.

Theorem 5.2 (Finite-depth susceptibility law)**.** *For every fixed depth d, there exists $u_*(d) > 0$ such that for all $|u| \leq u_*(d)$,*

$$m_d(u) = \chi_d\, u + O(u^3), \qquad \chi_d = \sum_{\ell=0}^{d} (b\rho)^\ell. \tag{5.4}$$

Equivalently, $\chi_d = m_d'(0)$ is the exact linear response of the root belief to an infinitesimal homogeneous repair bias.

Proof sketch. Appendix D proves that every finite-depth magnetization is real-analytic and odd near $u = 0$. Differentiating the exact recursion (5.1) at the origin yields

$$\chi_0 = 1, \qquad \chi_{d+1} = 1 + b\rho\, \chi_d,$$

whose solution is the geometric sum in (5.4). Oddness excludes the quadratic term, so the first nonlinear correction is cubic.

The quantity χ_d is the exact amplification factor by which recursive structure converts a weak local repair preference into a root-level decision bias. Its scaling immediately yields three regimes:

$$\begin{aligned} b\rho < 1 &\Rightarrow \chi_d = O(1), \\ b\rho = 1 &\Rightarrow \chi_d = d + 1, \\ b\rho > 1 &\Rightarrow \chi_d = \Theta\big((b\rho)^d\big). \end{aligned}$$

Thus the line $b\rho = 1$ is the exact finite-depth critical boundary separating bounded and exponentially amplified response.

To translate this susceptibility into a sentence-processing threshold, we now separate two sources of evidence. The clean ambiguous input contributes a *root-local* literal bias $u_0 < 0$, while teacher–student mismatch contributes a *distributed* repair bias Δ acting throughout the recursive structure. Define

$$G_d(\Delta) := \begin{cases} \Delta, & d = 0, \\ \Delta + b\, \operatorname{artanh}\big(\rho\, m_{d-1}(\Delta)\big), & d \geq 1, \end{cases} \tag{5.5}$$

so that

$$M_d(u_0, \Delta) = \tanh\big(u_0 + G_d(\Delta)\big). \tag{5.6}$$

Because $\tanh$ is strictly increasing, the Bayes decision at the root flips exactly when

$$u_0 + G_d(\Delta) = 0.$$

Theorem 5.3 (Hallucinated-repair threshold law)**.** *Assume $u_0 < 0$ and $\hat{J} \geq 0$. For sufficiently small $|u_0|$, there exists a unique critical mismatch*

$$\Delta_c(d) := \inf\{\Delta \geq 0 : M_d(u_0, \Delta) \geq 0\} \tag{5.7}$$

such that $M_d(u_0, \Delta_c(d)) = 0$, and

$$\Delta_c(d) = \frac{|u_0|}{\chi_d} + O\left(\frac{|u_0|^3}{\chi_d^3}\right). \tag{5.8}$$

Proof sketch. By Proposition 5.1, $G_d(\Delta)$ is strictly increasing, so the sign change is unique whenever it exists. Appendix E proves the small-mismatch expansion

$$G_d(\Delta) = \chi_d \, \Delta + O(\Delta^3),$$

which is the mixed-field counterpart of Theorem 5.2. The threshold equation $G_d(\Delta_c(d)) = |u_0|$ can therefore be inverted near the origin, giving (5.8).

Theorem 5.3 is the paper's central sentence-processing prediction. Recursive structure does not merely increase uncertainty; it *amplifies mismatch*. A clean ambiguous input may still locally favor the literal interpretation, but once χ_d is large enough, a mismatch of order χ_d^{-1} is sufficient to make the Bayes-optimal comprehender prefer a repaired interpretation that was never intended. In particular,

$$b\rho < 1 \Rightarrow \Delta_c(d) = \Theta(|u_0|),$$

$$b\rho = 1 \Rightarrow \Delta_c(d) = \Theta\left(\frac{|u_0|}{d+1}\right),$$

and

$$b\rho > 1 \Rightarrow \Delta_c(d) = \Theta\Big(|u_0| \, (b\rho)^{-d}\Big).$$

Thus subcritical trees require $O(1)$ mismatch, critical trees require $O(d^{-1})$ mismatch, and supercritical trees require exponentially small mismatch to trigger false repair.

Taken together, Theorems 5.2 and 5.3 yield a concrete theory of hallucinated repair: the Bayes-matched manifold of Section 3 is exactly calibrated, but off that manifold recursive syntax acts as a susceptibility amplifier. The next section tests these identities and threshold laws directly with exact synthetic inference and a symbolic natural-language evaluation.

6 Experimental Validation

We validate the theory in three stages: exact Bayes-matched identities, exact threshold laws, and a symbolic benchmark of omission-sensitive sentence-like inputs. Additional adversarial and structural robustness experiments are deferred to Appendix H.

6.1 Exact synthetic verification of the Nishimori identities

Figure 1 validates the two exact matched-manifold statements from Section 3 and then shows their controlled failure off the manifold. Teacher instances are generated exactly from the recursive repair grammar, and all student posterior quantities are computed by exact sum-product on the tree. The left and middle panels verify Corollary 3.2 and Corollary 3.3: expected truth overlap matches expected replica overlap, and Bayes risk matches posterior variance. The right panel shows the structured off-manifold failure mode: under pure field mismatch, the Nishimori gap opens immediately and separates sharply by regime, with the fastest growth in the supercritical case. Full numerical summaries appear in Table 4.

Table 1: Symbolic sanity-check benchmark for Section 6.3. 1296 recursive items spanning three function-word families, with 40 pseudo-observer responses per item sampled from the teacher posterior. The last column reports log loss on the top-Ξ quartile.

Model	Acc.	NLL	Brier	ECE	NLL (top-Ξ)
Literal-only	0.496	6.330	0.309	0.458	11.227
Uncoupled noisy channel	0.809	0.544	0.036	0.041	0.537
Recursive coupled noisy channel	**0.956**	**0.466**	**0.005**	**0.024**	**0.428**

6.2 Exact validation of susceptibility and threshold laws

We next test the two central quantitative predictions of Section 5: the finite-depth susceptibility law (Theorem 5.2) and the hallucinated-repair threshold law (Theorem 5.3). The susceptibility $\hat{\chi}_d$ is estimated independently from exact magnetization curves, and the threshold $\hat{\Delta}_c$ is obtained independently by monotone bisection on the exact mixed-field response. Figure 2 shows that both laws hold at numerical precision: the left panel confirms the geometric susceptibility law, the middle panel validates the threshold collapse, and the right panel makes the subcritical/critical/supercritical scaling distinction explicit. Full numerical summaries are reported in Table 5.

6.3 Symbolic natural-language sanity check

To test whether the same mechanism remains visible on interpretable sentence-like inputs, we construct a symbolic benchmark of omission-sensitive complementizer, relative-pronoun, and infinitival-marker items. The benchmark contains 1296 human-readable recursive items with teacher-sampled pseudo-observer responses and is evaluated under a literal-only baseline, an uncoupled noisy-channel baseline, and the recursive coupled model.

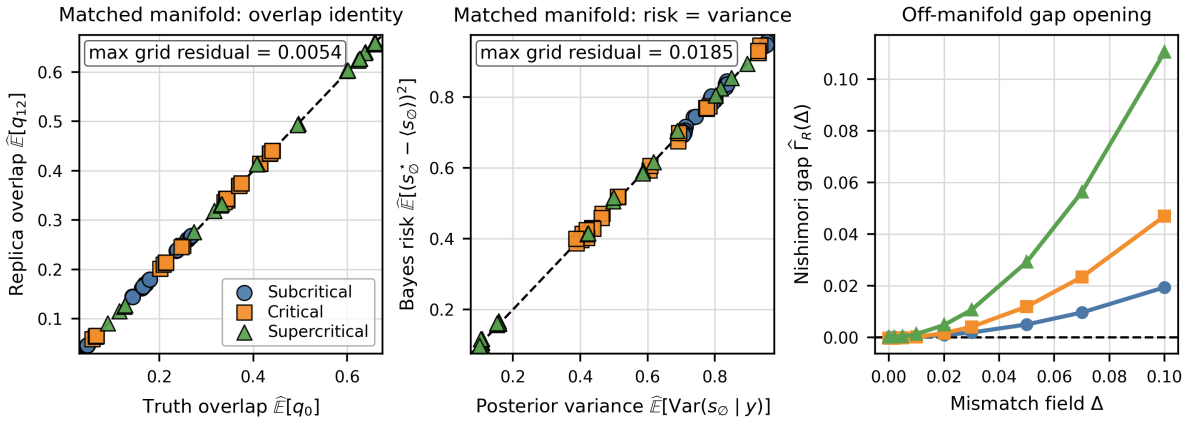

Figure 1: **Exact synthetic verification of the Nishimori identities.** Left: truth overlap versus replica overlap on the Bayes-matched manifold. Middle: Bayes risk versus posterior variance. Right: off-manifold opening of the Nishimori gap under pure field mismatch.

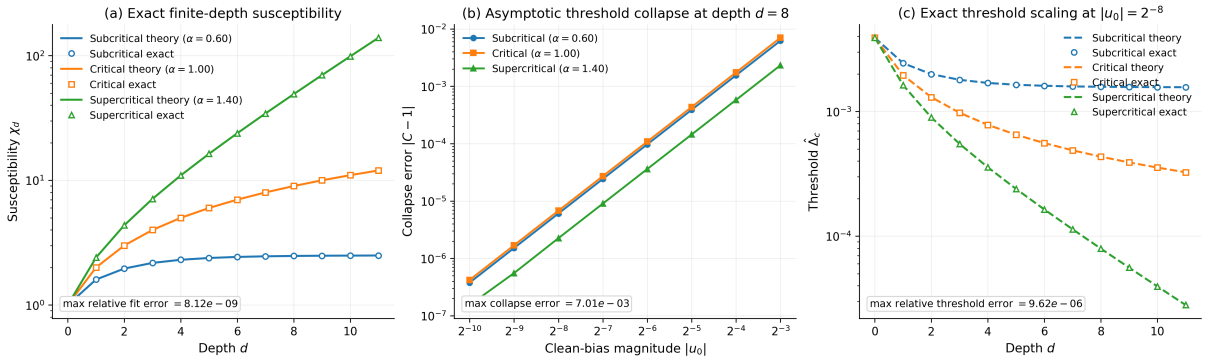

Figure 2: **Exact validation of the susceptibility and threshold laws.** Left: independently estimated susceptibility versus the closed-form law. Middle: threshold-collapse statistic $C = \hat{\Delta}_c \chi_d / |u_0|$ at fixed depth. Right: exact threshold scaling across depth in the subcritical, critical, and supercritical regimes.

Table 1 shows that the recursive coupled model substantially outperforms both baselines in accuracy, log loss, Brier score, and calibration error. The gain is largest on the top-Ξ quartile, where $\Xi = \chi_d \times$ omitted fraction serves as the benchmark-level recursive-drive statistic. This is exactly the regime in which the theory predicts recursive amplification should matter most. Figure 3 and the additional robustness results appear in Appendix H.

7 Related Work

Our closest point of contact is noisy-channel sentence processing, from Levy (2008) and Levy (2011) through later work on rational correction, noisy context, adaptation to local noise ecologies, and algorithmic constraints on approximate inference (Gibson et al., 2013; Futrell and Levy, 2017; Ryskin et al., 2018; Chen et al., 2023; Clark et al., 2025). Our work differs in mathematical object: rather than proposing another processing model or approximation scheme, we identify a Bayes-matched manifold for sentence repair, prove exact overlap and variance identities on that manifold, and derive a finite-depth threshold law for off-manifold hallucinated repair using Nishimori structure from Bayesian statistical physics (Nishimori, 2001).

8 Conclusion

We introduced a Bayes-matched Nishimori manifold for noisy-channel sentence processing, proved exact calibration identities on that manifold, and derived an inverse-susceptibility threshold law for hallucinated repair in a recursive repair grammar; exact synthetic experiments and a symbolic benchmark support the same mechanism.

Limitations

The exact theorems are proved for controlled teacher–student families and, in the critical analy-

sis, for recursive repair grammars whose posterior reduces to a tree-structured spin system; the main-text experiments are exact synthetic validations plus a symbolic benchmark with teacher-sampled responses, so they isolate Bayes matching, mismatch, and recursive amplification but are not direct evidence about human implementation.

References

Sihan Chen, Sarah Nathaniel, Rachel Ryskin, and Edward Gibson. 2023. The effect of context on noisy-channel sentence comprehension. *Cognition*, 238:105503.

Thomas Hikaru Clark, Jacob Hoover Vigly, Edward Gibson, and Roger P. Levy. 2025. Resource-rational noisy-channel language processing: Testing the effect of algorithmic constraints on inferences. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23648–23661, Suzhou, China. Association for Computational Linguistics.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain. Association for Computational Linguistics.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Roger Levy. 2008. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Roger Levy. 2011. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, Portland, Oregon, USA. Association for Computational Linguistics.

Hidetoshi Nishimori. 2001. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press.

Rachel Ryskin, Richard Futrell, Swathi Kiran, and Edward Gibson. 2018. Comprehenders model the nature of noise in the environment. *Cognition*, 181:141–150.

A Proof of the Sentence Nishimori Identity and Corollaries

This appendix gives complete proofs of Theorem 3.1, Corollary 3.2, and Corollary 3.3. The argument is the exact Bayes-matched exchangeability mechanism characteristic of Nishimori theory (Nishimori, 2001), specialized here to the noisy-channel sentence-processing model introduced in Section 2 and to the exact posterior formulation used in noisy-channel accounts of sentence comprehension (Levy, 2008, 2011). No asymptotic approximation, replica trick, or algorithmic approximation is used in this appendix.

Setup. Throughout Appendix A, we work on the matched manifold $\mathcal{N}$, so that the student and teacher parameters agree:

$$(\hat{\theta}, \hat{\eta}) = (\theta, \eta).$$

Accordingly, we abbreviate the matched posterior by

$$\begin{aligned} \Pi(z \mid x, c) &:= \Pi_{\theta,\eta}(z \mid x, c) \\ &= \frac{P_\theta(z \mid c)\, Q_\eta(x \mid z, c)}{\mathsf{Z}_{\theta,\eta}(x, c)}, \\ \mathsf{Z}_{\theta,\eta}(x, c) &:= \sum_{z' \in \mathcal{Z}} P_\theta(z' \mid c)\, Q_\eta(x \mid z', c). \end{aligned} \tag{A.1}$$

Assumption 2.1 guarantees that $\Pi(\cdot \mid x, c)$ is a proper probability distribution for every (x, c) in the support of (X, C).

Because $\mathcal{Z}$ is finite or countable, $(x, c) \mapsto \Pi(\cdot \mid x, c)$ is a probability kernel on a discrete latent space. Hence, on an auxiliary extension of the base probability space, one may realize an infinite sequence of posterior replicas $(Z^{(k)})_{k \geq 1}$ such that, conditional on (X, C), the replicas are i.i.d. with common law $\Pi(\cdot \mid X, C)$ and are conditionally independent of $Z^\star$ given (X, C). This is the probability space underlying all expectations below.

Lemma A.1 (Conditional posterior law of the planted state)**.** *On the matched manifold $\mathcal{N}$, for every $z \in \mathcal{Z}$ and every (x, c) in the support of (X, C),*

$$\mathbb{P}(Z^\star = z \mid X = x, C = c) = \Pi(z \mid x, c). \tag{A.2}$$

Equivalently, conditional on (X, C), the planted latent state $Z^\star$ has the same law as a posterior replica.

Proof. Fix (x, c) in the support of (X, C). By Bayes' rule and the teacher model in (2.1),

$$\begin{aligned}
\mathsf{Z}_{\theta,\eta}(x, c) &:= \sum_{z' \in \mathcal{Z}} P_\theta(z' \mid c)\, Q_\eta(x \mid z', c), \\
\Pi(z \mid x, c) &= \mathbb{P}(Z^\star = z \mid X = x, C = c) \\
&= \frac{\mathbb{P}(Z^\star = z, X = x \mid C = c)}{\mathbb{P}(X = x \mid C = c)} \\
&= \frac{P_\theta(z \mid c)\, Q_\eta(x \mid z, c)}{\mathsf{Z}_{\theta,\eta}(x, c)}. \qquad \text{(A.3)}
\end{aligned}$$

where the denominator is finite and strictly positive by Assumption 2.1. This is exactly (A.2). □

Lemma A.2 (Conditional exchangeability of the planted state and replicas)**.** *Fix $k \geq 1$. On $\mathcal{N}$, conditional on (X, C), the random variables*

$$Z^\star,\ Z^{(1)},\ Z^{(2)},\ \ldots,\ Z^{(k)}$$

are i.i.d. with common law $\Pi(\cdot \mid X, C)$. In particular, for every $z_0, z_1, \ldots, z_k \in \mathcal{Z}$,

$$\begin{aligned}
&\mathbb{P}\Big(Z^\star = z_0,\ Z^{(1)} = z_1,\ \ldots,\ Z^{(k)} = z_k \mid X, C \Big) \\
&\qquad = \prod_{j=0}^{k} \Pi(z_j \mid X, C). \qquad \text{(A.4)}
\end{aligned}$$

where z_0 corresponds to $Z^\star$.

Proof. By Lemma A.1, the planted state $Z^\star$ has conditional law $\Pi(\cdot \mid X, C)$. By construction of the posterior replicas, $(Z^{(j)})_{j \geq 1}$ are conditionally i.i.d. with the same law $\Pi(\cdot \mid X, C)$ and are conditionally independent of $Z^\star$ given (X, C). Multiplying the corresponding conditional masses yields (A.4). □

Proof of Theorem 3.1. Let $k \geq 1$, and let f be any bounded measurable observable of

$$(Z^\star, Z^{(1)}, \ldots, Z^{(k)}, X, C).$$

Since $\mathcal{Z}$ is discrete, measurability is immediate. By the tower property,

$$\begin{aligned}
F_k &:= f\Big(Z^\star, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big), \\
\mathbb{E}[F_k] &= \mathbb{E}[\mathbb{E}[F_k \mid X, C]]. \qquad \text{(A.5)}
\end{aligned}$$

Now fix (x, c) in the support of (X, C). By Lemma A.2,

$$\begin{aligned}
&\mathbb{E}\Big[f\Big(Z^\star, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big) \,\Big|\, X = x, C = c\Big] \\
&= \sum_{z_0, \ldots, z_k \in \mathcal{Z}} f(z_0, z_1, \ldots, z_k, x, c) \\
&\qquad \times \prod_{j=0}^{k} \Pi(z_j \mid x, c). \qquad \text{(A.6)}
\end{aligned}$$

Introduce one more dummy summation variable z_{k+1}. Since all factors are identical copies of the same posterior mass function, merely renaming the dummy variable z_0 as z_{k+1} gives

$$\begin{aligned}
&\sum_{z_0, \ldots, z_k \in \mathcal{Z}} f(z_0, z_1, \ldots, z_k, x, c) \prod_{j=0}^{k} \Pi(z_j \mid x, c) \\
&= \sum_{z_1, \ldots, z_{k+1} \in \mathcal{Z}} f(z_{k+1}, z_1, \ldots, z_k, x, c) \\
&\qquad \times \prod_{j=1}^{k+1} \Pi(z_j \mid x, c). \qquad \text{(A.7)}
\end{aligned}$$

Applying Lemma A.2 again, the right-hand side of (A.7) is exactly

$$\mathbb{E}\Big[f\Big(Z^{(k+1)}, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big) \Big| X = x, C = c\Big]. \qquad \text{(A.8)}$$

Combining (A.6)–(A.8) and then integrating over (X, C) via (A.5) yields

$$\begin{aligned}
F_\star &:= f\Big(Z^\star, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big), \\
F_{k+1} &:= f\Big(Z^{(k+1)}, Z^{(1)}, \ldots, Z^{(k)}, X, C\Big), \\
\mathbb{E}[F_\star] &= \mathbb{E}[F_{k+1}].
\end{aligned}$$

which proves Theorem 3.1. □

Proof of Corollary 3.2. Let $R : \mathcal{Z} \times \mathcal{Z} \to [-1, 1]$ be any bounded overlap observable. Apply Theorem 3.1 with $k = 1$ and

$$f(z_0, z_1, x, c) := R(z_0, z_1).$$

Then

$$\mathbb{E}\Big[R(Z^\star, Z^{(1)})\Big] = \mathbb{E}\Big[R(Z^{(2)}, Z^{(1)})\Big]. \qquad \text{(A.9)}$$

Since the posterior replicas are conditionally i.i.d. given (X, C), the pair $(Z^{(1)}, Z^{(2)})$ is exchangeable, hence

$$\mathbb{E}\Big[R(Z^{(2)}, Z^{(1)})\Big] = \mathbb{E}\Big[R(Z^{(1)}, Z^{(2)})\Big]. \qquad \text{(A.10)}$$

Combining (A.9) and (A.10) gives

$$\mathbb{E}\Big[R(Z^\star, Z^{(1)})\Big] = \mathbb{E}\Big[R(Z^{(1)}, Z^{(2)})\Big].$$

By the definition of the Nishimori gap in (2.4),

$$\Gamma_R = \mathbb{E}\Big[R(Z^{(1)}, Z^{(2)}) - R(Z^\star, Z^{(1)})\Big] = 0.$$

This proves Corollary 3.2. □

Proof of Corollary 3.3. Let $\phi : \mathcal{Z} \to \mathbb{R}$ be square-integrable. Define the posterior mean

$$m_\phi(X, C) := \langle \phi(Z) \rangle = \sum_{z \in \mathcal{Z}} \phi(z)\, \Pi(z \mid X, C).$$

By square integrability and the tower property,

$$\Delta_\phi := \phi(Z^\star) - m_\phi(X, C),$$
$$\mathbb{E}\Big[\big(\phi(Z^\star) - \langle \phi(Z) \rangle\big)^2\Big] = \mathbb{E}\big[\mathbb{E}[\Delta_\phi^2 \mid X, C]\big]. \tag{A.11}$$

Conditional on (X, C), Lemma A.1 implies that $Z^\star$ has law $\Pi(\cdot \mid X, C)$. Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[\big(\phi(Z^\star) - m_\phi(X, C)\big)^2 \,\Big|\, X, C\Big] \\
&= \sum_{z \in \mathcal{Z}} \big(\phi(z) - m_\phi(X, C)\big)^2 \Pi(z \mid X, C) \\
&= \sum_{z \in \mathcal{Z}} \phi(z)^2 \Pi(z \mid X, C) \\
&\quad - \left(\sum_{z \in \mathcal{Z}} \phi(z) \Pi(z \mid X, C)\right)^2 \\
&= \big\langle \phi(Z)^2 \big\rangle - \langle \phi(Z) \rangle^2 .
\end{aligned} \tag{A.12}$$

Substituting (A.12) into (A.11) yields

$$\mathbb{E}\Big[\big(\phi(Z^\star) - \langle \phi(Z) \rangle\big)^2\Big] = \mathbb{E}\Big[\big\langle \phi(Z)^2 \big\rangle - \langle \phi(Z) \rangle^2\Big],$$

which is exactly Corollary 3.3. □

Interpretation. Theorem 3.1 and Corollaries 3.2–3.3 formalize the privileged status of the Bayes-matched manifold $\mathcal{N}$. On $\mathcal{N}$, internal posterior self-consistency is exactly calibrated to truth consistency, and posterior uncertainty is exactly equal to Bayes risk. This is the baseline from which all later off-manifold phenomena—in particular, hallucinated repair under teacher–student mismatch—must be measured.

B Explicit Construction of the Recursive Repair Grammar

This appendix gives a fully explicit construction of the finite weighted grammar underlying Section 4 and proves the exact posterior mapping stated there. The construction is intentionally concrete: rather than postulating an Ising posterior directly, we show that it arises as the exact latent posterior of a recursive noisy-channel grammar over strings. This places the model in the weighted-grammar noisy-channel tradition of Levy (2008, 2011), while departing from prior work by introducing *recursive couplings* between repair hypotheses.

B.1 Underlying recursive skeleton and deterministic linearization

Fix an integer branching factor $b \geq 1$ and depth $d \geq 0$. Let

$$V_{b,d} := \bigcup_{t=0}^{d} [b]^t$$

be the vertex set of the rooted ordered b-ary tree of depth d, where $[b]^t$ denotes the set of all length-t sequences over $\{1, \ldots, b\}$. The root is the empty sequence $\varnothing$, and the edge set is

$$E_{b,d} := \{(v, vi) : v \in [b]^t,\ 0 \leq t < d,\ 1 \leq i \leq b\}.$$

For $v \in V_{b,d}$, let $|v|$ denote its depth.

A latent repair assignment is a spin configuration

$$s = (s_v)_{v \in V_{b,d}} \in \{-1, +1\}^{V_{b,d}},$$

where $s_v = -1$ denotes the literal hypothesis at site v and $s_v = +1$ denotes the repaired hypothesis. Each node also carries a local observed cue

$$y_v \in \mathcal{Y},$$

where $\mathcal{Y}$ is a finite cue alphabet. In the main text, y_v will be interpreted as a repair-relevant local signal, such as the presence or absence of a low-information structural marker; for the present appendix only finiteness of $\mathcal{Y}$ matters.

To turn the tree-indexed cue field $y = (y_v)_{v \in V_{b,d}}$ into an observed string, introduce distinct depth-indexed delimiter tokens

$$\{\ell_t, r_t : 0 \leq t \leq d\}$$

that are disjoint from $\mathcal{Y}$. Define the deterministic linearization map

$$\begin{aligned}
&\mathrm{Lin}_{b,d} : \mathcal{Y}^{V_{b,d}} \to \Sigma^*, \\
&\qquad \Sigma := \mathcal{Y} \cup \{\ell_t, r_t : 0 \leq t \leq d\}.
\end{aligned}$$

recursively by

$$\mathrm{Lin}_{b,d}(y; v) = \ell_{|v|}\, y_v\, \mathrm{Lin}_{b,d}(y; v1) \cdots \mathrm{Lin}_{b,d}(y; vb)\, r_{|v|} \tag{B.1}$$

for $|v| < d$, and

$$\mathrm{Lin}_{b,d}(y; v) = \ell_d\, y_v\, r_d \tag{B.2}$$

for $|v| = d$. The observed string is

$$x = \mathrm{Lin}_{b,d}(y; \varnothing).$$

Lemma B.1 (Injective linearization)**.** *The map* $\mathrm{Lin}_{b,d}$ *is injective. Consequently, conditioning on the observed string* x *is equivalent to conditioning on the cue field* y*.*

Proof. Because the delimiters ℓ_t, r_t are depth-specific and distinct from $\mathcal{Y}$, the recursive decomposition (B.1)–(B.2) uniquely determines the ordered tree structure and the depth of every node. One can therefore recover each cue y_v by recursively parsing the unique symbol immediately following $\ell_{|v|}$ in the corresponding subtree span. This reconstructs y uniquely from x. □

Lemma B.1 is an analytic convenience: it ensures that the observed string carries exactly the same information as the node-indexed cue field. The depth-indexed delimiters are not meant as a linguistic claim; they merely make the controlled recursive family exactly solvable and unambiguous. In the natural-language sanity check, these analytic delimiters will be replaced by ordinary clause templates and function-word environments.

B.2 Weighted grammar construction

We now define a finite weighted context-free grammar $\mathcal{G}_{b,d}(c)$ for each discourse context c.

Definition B.2 (Recursive repair grammar)**.** Fix teacher parameters $\theta = (J, h)$, where $J \in \mathbb{R}$ is a coupling strength and $h : \mathcal{C} \to \mathbb{R}$ is a context field. Let $q_\eta(\cdot \mid \sigma, c)$ be a probability mass function on $\mathcal{Y}$ for each $\sigma \in \{-1, +1\}$ and $c \in \mathcal{C}$. The weighted grammar $\mathcal{G}_{b,d}(c)$ has:

- start symbol S;
- nonterminals
$$\mathcal{N} = \{S_t^+, S_t^- : 0 \le t \le d\};$$
- terminal alphabet $\Sigma = \mathcal{Y} \cup \{\ell_t, r_t : 0 \le t \le d\}$;
- start rules
$$S \to S_0^+, \qquad S \to S_0^-,$$
each with weight 1;
- for each depth $0 \le t < d$, sign $\sigma \in \{-1, +1\}$, child-sign tuple $\tau_{1:b} \in \{-1, +1\}^b$, and cue $y \in \mathcal{Y}$, an internal-node rule
$$S_t^\sigma \to \ell_t\, y\, S_{t+1}^{\tau_1} \cdots S_{t+1}^{\tau_b}\, r_t \tag{B.3}$$
with weight
$$\omega_t(\sigma, y, \tau_{1:b} \mid c) := \exp\left(h(c)\sigma + J \sum_{i=1}^{b} \sigma\tau_i\right) \times q_\eta(y \mid \sigma, c). \tag{B.4}$$
- for each leaf sign $\sigma \in \{-1, +1\}$ and cue $y \in \mathcal{Y}$, a leaf rule
$$S_d^\sigma \to \ell_d\, y\, r_d \tag{B.5}$$
with weight
$$\omega_d(\sigma, y \mid c) := \exp\big(h(c)\sigma\big)\, q_\eta(y \mid \sigma, c). \tag{B.6}$$

The grammar is finite because both $V_{b,d}$ and $\mathcal{Y}$ are finite. Since all rule weights are strictly positive whenever $q_\eta(y \mid \sigma, c) > 0$, every complete derivation has positive weight.

Remark B.3. For clarity we use a depth-independent context field $h(c)$. Every derivation below goes through unchanged if $h(c)$ is replaced by a depth-dependent field $h_{|v|}(c)$; the main text uses the simpler homogeneous form because it already suffices to generate the susceptibility transition of interest.

Proposition B.4 (Unique derivation for a fixed latent assignment and cue field)**.** *For every pair* $(s, y) \in \{-1, +1\}^{V_{b,d}} \times \mathcal{Y}^{V_{b,d}}$*, there exists exactly one complete derivation* $D(s, y)$ *in* $\mathcal{G}_{b,d}(c)$ *whose hidden nonterminal state at node* v *is* $S_{|v|}^{s_v}$ *and whose emitted cue at* v *is* y_v*. The terminal yield of* $D(s, y)$ *is* $x = \mathrm{Lin}_{b,d}(y; \varnothing)$*.*

Proof. Proceed by induction on the depth of a subtree. At a leaf v with $|v| = d$, the sign s_v and cue y_v uniquely determine the leaf rule (B.5). Now suppose $|v| = t < d$. Given s_v, y_v, and the child signs $s_{v1}, \ldots, s_{vb}$, exactly one internal-node rule of the form (B.3) is compatible with that tuple. By

the induction hypothesis, each child subtree has a unique derivation. Therefore the full subtree rooted at v also has a unique derivation. Applying this at the root yields uniqueness of $D(s, y)$, and the terminal yield follows immediately from the recursive form of the rules. □

Proposition B.5 (Exact factorization of derivation weights)**.** *Let $D(s, y)$ be the unique derivation associated with (s, y). Then its total weight factorizes as*

$$W_c\big(D(s,y)\big) = \exp\left(h(c)\sum_{v\in V_{b,d}} s_v + J\sum_{(u,v)\in E_{b,d}} s_u s_v\right) \times \prod_{v\in V_{b,d}} q_\eta(y_v \mid s_v, c). \tag{B.7}$$

Proof. By definition, the weight of a complete derivation is the product of its rule weights. Every node v contributes exactly one factor $\exp(h(c)s_v)$ from either (B.4) or (B.6). Every edge (u, v) contributes exactly one factor $\exp(Js_u s_v)$, appearing in the expansion rule of the parent u. Every node contributes exactly one channel factor $q_\eta(y_v \mid s_v, c)$. Multiplying these contributions over all nodes and edges gives (B.7). □

B.3 Teacher prior and observation channel induced by the grammar

We now extract the probabilistic teacher model (P_θ, Q_η) from the weighted grammar.

Definition B.6 (Teacher normalizer)**.** Define

$$\begin{aligned} H_c(s) &:= h(c)\sum_v s_v + J\sum_{(u,v)\in E_{b,d}} s_u s_v, \\ Z_\theta(c) &:= \sum_{s\in\{-1,+1\}^{V_{b,d}}} \exp\big(H_c(s)\big). \end{aligned} \tag{B.8}$$

Since $V_{b,d}$ is finite, $Z_\theta(c) \in (0, \infty)$.

Lemma B.7 (Teacher prior and local channel)**.** *The weighted grammar $\mathcal{G}_{b,d}(c)$ induces the following teacher prior over latent repair assignments:*

$$P_\theta(s \mid c) = \frac{1}{Z_\theta(c)} \exp\left(h(c)\sum_v s_v + J\sum_{(u,v)\in E_{b,d}} s_u s_v\right). \tag{B.9}$$

Conditional on s, the cue field y factorizes as

$$Q_\eta(y \mid s, c) = \prod_{v\in V_{b,d}} q_\eta(y_v \mid s_v, c). \tag{B.10}$$

Equivalently, the channel over observed strings $x \in \Sigma^*$ is

$$\begin{aligned} I_x(y) &:= \mathbb{I}\{x = \mathrm{Lin}_{b,d}(y; \varnothing)\}, \\ Q_\eta(x \mid s, c) &= \sum_{y\in\mathcal{Y}^{V_{b,d}}} I_x(y) \prod_{v\in V_{b,d}} q_\eta(y_v \mid s_v, c). \end{aligned} \tag{B.11}$$

Proof. By Proposition B.5, the unnormalized joint weight of (s, y) is exactly

$$\exp\left(h(c)\sum_v s_v + J\sum_{(u,v)} s_u s_v\right) \times \prod_v q_\eta(y_v \mid s_v, c).$$

Summing over all cue assignments y yields

$$\begin{aligned} S_c(s) &:= \sum_{y\in\mathcal{Y}^{V_{b,d}}} W_c\big(D(s,y)\big), \\ H_c(s) &:= h(c)\sum_v s_v + J\sum_{(u,v)} s_u s_v, \\ S_c(s) &= e^{H_c(s)} \prod_v \sum_{a\in\mathcal{Y}} q_\eta(a \mid s_v, c). \end{aligned}$$

Because $q_\eta(\cdot \mid s_v, c)$ is a probability mass function, the product of sums is 1, giving (B.9) after normalization by $Z_\theta(c)$. Equation (B.10) is simply the conditional factorization of the cue weights, and (B.11) follows from the deterministic linearization map. □

Lemma B.7 is the key bridge back to Section 2: in the recursive grammar family, the latent sentence state is the repair assignment s, and the observed sentence is the linearized string $x = \mathrm{Lin}_{b,d}(y; \varnothing)$. Because linearization is injective (Lemma B.1), inference conditioned on x is equivalent to inference conditioned on the local cue field y.

B.4 Proof of the exact posterior mapping

We now prove the exact posterior representation stated as Theorem 4.1 in Section 4.

Assumption B.8 (Strictly positive student channel)**.** For every cue $y \in \mathcal{Y}$, state $\sigma \in \{-1, +1\}$, and context $c \in \mathcal{C}$,

$$\hat{q}_{\hat{\eta}}(y \mid \sigma, c) > 0.$$

Assumption B.8 is used only to write the local likelihood contribution as a finite log-odds field. If zeros are permitted, the same derivation goes through with $+\infty$ or $-\infty$ fields on the support boundary.

Proof of Theorem 4.1. Let the student parameters be $\hat{\theta} = (\hat{J}, \hat{h})$ and $\hat{\eta}$, where $\hat{h} : \mathcal{C} \to \mathbb{R}$ and $\hat{q}_{\hat{\eta}}(\cdot \mid \sigma, c)$ is a positive probability mass function on $\mathcal{Y}$. By Lemma B.1, it is enough to compute the student posterior conditioned on the cue field y rather than on the corresponding string x.

Using the teacher–student form (2.2) specialized to the grammar family, we obtain

$$\hat{H}_c(s) := \hat{h}(c) \sum_{v} s_v + \hat{J} \sum_{(u,v)\in E_{b,d}} s_u s_v,$$

$$\Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) \propto \exp\big(\hat{H}_c(s)\big) \prod_{v\in V_{b,d}} \hat{q}_{\hat{\eta}}(y_v \mid s_v, c). \tag{B.12}$$

For each fixed $y \in \mathcal{Y}$ and $c \in \mathcal{C}$, define

$$a_{\hat{\eta}}(y, c) := \frac{1}{2} \log\Big(\hat{q}_{\hat{\eta}}(y \mid +1, c)\, \hat{q}_{\hat{\eta}}(y \mid -1, c)\Big), \tag{B.13}$$

$$b_{\hat{\eta}}(y, c) := \frac{1}{2} \log \frac{\hat{q}_{\hat{\eta}}(y \mid +1, c)}{\hat{q}_{\hat{\eta}}(y \mid -1, c)}. \tag{B.14}$$

Because $s_v \in \{-1, +1\}$, every local log-likelihood can be written exactly as

$$\log \hat{q}_{\hat{\eta}}(y_v \mid s_v, c) = a_{\hat{\eta}}(y_v, c) + b_{\hat{\eta}}(y_v, c)\, s_v. \tag{B.15}$$

Substituting (B.15) into (B.12) yields

$$\begin{aligned} \lambda_v(y, c) &:= \hat{h}(c) + b_{\hat{\eta}}(y_v, c), \\ H_{y,c}(s) &:= \hat{J} \sum_{(u,v)\in E_{b,d}} s_u s_v + \sum_{v\in V_{b,d}} \lambda_v(y, c)\, s_v, \\ A(y, c) &:= \sum_{v\in V_{b,d}} a_{\hat{\eta}}(y_v, c), \\ \Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) &\propto \exp\big(H_{y,c}(s)\big) \exp\big(A(y, c)\big). \end{aligned} \tag{B.16}$$

The final exponential factor in (B.16) is independent of s and is therefore absorbed into the normalization constant. Consequently,

$$H_{y,c}(s) := \hat{J} \sum_{(u,v)\in E_{b,d}} s_u s_v + \sum_{v\in V_{b,d}} u_{\hat{\eta}}(y_v, c)\, s_v, \quad \Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) \propto \exp\big(H_{y,c}(s)\big). \tag{B.17}$$

where the effective local field is

$$u_{\hat{\eta}}(y_v, c) := \hat{h}(c) + \frac{1}{2} \log \frac{\hat{q}_{\hat{\eta}}(y_v \mid +1, c)}{\hat{q}_{\hat{\eta}}(y_v \mid -1, c)}. \tag{B.18}$$

Equation (B.17) is exactly the random-field Ising representation claimed in Theorem 4.1. □

Remark B.9 (Why the construction matters). Theorem 4.1 is not obtained by analogy to statistical physics; it is the exact latent posterior of a finite weighted grammar over strings. The recursive couplings $J s_u s_v$ arise from grammar-mediated dependencies between repair choices across clause levels, while the random field $u_{\hat{\eta}}(y_v, c)$ is the student's local log-odds in favor of repair at each site. This exact grammar-to-Ising correspondence is what makes the later susceptibility analysis mathematically sharp rather than metaphorical.

C Exact Cavity / Belief-Propagation Recursion

This appendix derives the exact cavity recursion for the recursive repair grammar introduced in Section 4. The derivation is exact because the posterior factor graph is a tree: no Bethe approximation, mean-field closure, or algorithmic truncation is used. The cavity equations obtained here are the sole input to the susceptibility and threshold analysis in Appendix D and Appendix E.

C.1 Tree-structured posterior and constrained subtree partition functions

Fix a finite rooted tree $T = (V, E)$ with root r. In the recursive repair family of Section 4, T will later be specialized to the regular tree $T_{b,d}$, but the cavity derivation does not require regularity. Let

$$\mathrm{ch}(v) := \{w \in V : (v, w) \in E\}$$

denote the children of a vertex v, and for every non-root vertex v, let $p(v)$ denote its unique parent.

After Theorem 4.1, the student posterior on a fixed observed cue field y and context c takes the exact Gibbs form

$$\begin{aligned} u_v &:= u_{\hat{\eta}}(y_v, c), \\ H_{y,c}(s) &:= \hat{J} \sum_{(u,v)\in E} s_u s_v + \sum_{v\in V} u_v s_v, \\ \Pi_{\hat{\theta},\hat{\eta}}(s \mid y, c) &\propto \exp\big(H_{y,c}(s)\big). \end{aligned} \tag{C.1}$$

where $s_v \in \{-1, +1\}$ and $\hat{J} \in \mathbb{R}$ is the student's coupling parameter. Throughout this appendix we assume $\hat{J}$ is finite and define

$$\rho := \tanh(\hat{J}), \qquad |\rho| < 1. \tag{C.2}$$

For each directed edge $v \to p(v)$, let T_v denote the descendant subtree rooted at v, with the edge to its parent removed. Write V_v and E_v for its

vertex and edge sets. For $\sigma \in \{-1,+1\}$, define the *constrained subtree partition functions*

$$H_v(s) := \hat{J} \sum_{(a,b)\in E_v} s_a s_b + \sum_{a\in V_v} u_a s_a,$$
$$Z^{\sigma}_{v\to p(v)} := \sum_{\substack{s_{V_v}\in\{-1,+1\}^{V_v}\\ s_v=\sigma}} \exp\big(H_v(s)\big). \tag{C.3}$$

These partition functions exclude the interaction between v and its parent. They are the natural cavity objects because, on a tree, conditioning on s_v disconnects the subtrees rooted at the children of v.

Proposition C.1 (Exact subtree recursion)**.** *For every non-root vertex v and $\sigma \in \{-1,+1\}$,*

$$Z^{\sigma}_{v\to p(v)} = e^{u_v\sigma} \prod_{w\in\mathrm{ch}(v)} \left(\sum_{\tau\in\{-1,+1\}} e^{\hat{J}\sigma\tau}\, Z^{\tau}_{w\to v} \right). \tag{C.4}$$

For a leaf v, the empty product convention gives

$$Z^{\sigma}_{v\to p(v)} = e^{u_v\sigma}. \tag{C.5}$$

Proof. Fix $v \neq r$ and condition on $s_v = \sigma$. Since T is a tree, once the parent edge is removed, the descendants of distinct children $w \in \mathrm{ch}(v)$ lie in disjoint subtrees that do not interact with one another except through s_v. Therefore the Gibbs weight in (C.3) factorizes over child subtrees:

$$H_w(s) := \hat{J} \sum_{(a,b)\in E_w} s_a s_b + \sum_{a\in V_w} u_a s_a,$$
$$\hat{J} \sum_{(a,b)\in E_v} s_a s_b + \sum_{a\in V_v} u_a s_a = u_v\sigma + \sum_{w\in\mathrm{ch}(v)} \big(\hat{J}\sigma s_w + H_w(s)\big).$$

Summing over all assignments in T_v with root spin fixed to σ thus factorizes into a product of child-subtree sums, yielding (C.4). When v is a leaf, there are no children and the product is empty, giving (C.5). □

C.2 Exact ratio and magnetization messages

The subtree partition functions are positive, so one may equivalently work with likelihood ratios or magnetization messages.

Definition C.2 (Cavity ratio and cavity magnetization)**.** For every directed edge $v \to p(v)$, define

$$R_{v\to p(v)} := \frac{Z^{+}_{v\to p(v)}}{Z^{-}_{v\to p(v)}}, \tag{C.6}$$

and

$$m_{v\to p(v)} := \frac{Z^{+}_{v\to p(v)} - Z^{-}_{v\to p(v)}}{Z^{+}_{v\to p(v)} + Z^{-}_{v\to p(v)}} = \frac{R_{v\to p(v)} - 1}{R_{v\to p(v)} + 1}. \tag{C.7}$$

Proposition C.3 (Exact ratio recursion)**.** *For every non-root vertex v,*

$$R_{v\to p(v)} = e^{2u_v} \prod_{w\in\mathrm{ch}(v)} \frac{e^{\hat{J}} R_{w\to v} + e^{-\hat{J}}}{e^{-\hat{J}} R_{w\to v} + e^{\hat{J}}}. \tag{C.8}$$

Proof. Divide the $\sigma = +1$ version of (C.4) by the $\sigma = -1$ version. This gives

$$R_{v\to p(v)} = e^{2u_v} \prod_{w\in\mathrm{ch}(v)} \frac{e^{\hat{J}} Z^{+}_{w\to v} + e^{-\hat{J}} Z^{-}_{w\to v}}{e^{-\hat{J}} Z^{+}_{w\to v} + e^{\hat{J}} Z^{-}_{w\to v}}.$$

Factoring out $Z^{-}_{w\to v}$ from numerator and denominator of each child term yields (C.8). □

The ratio recursion becomes especially transparent in magnetization variables.

Proposition C.4 (Exact cavity magnetization recursion)**.** *For every non-root vertex v,*

$$A_v := u_v + \sum_{w\in\mathrm{ch}(v)} \mathrm{artanh}\big(\rho\, m_{w\to v}\big),$$
$$m_{v\to p(v)} = \tanh(A_v). \tag{C.9}$$

Equivalently,

$$A_v := e^{2u_v} \prod_{w\in\mathrm{ch}(v)} \big(1 + \rho\, m_{w\to v}\big),$$
$$B_v := \prod_{w\in\mathrm{ch}(v)} \big(1 - \rho\, m_{w\to v}\big), \tag{C.10}$$
$$m_{v\to p(v)} = \frac{A_v - B_v}{A_v + B_v}.$$

Proof. Starting from (C.8), use (C.7) to write

$$R_{w\to v} = \frac{1 + m_{w\to v}}{1 - m_{w\to v}}.$$

A direct computation gives

$$A_{w\to v} := \frac{e^{\hat{J}} R_{w\to v} + e^{-\hat{J}}}{e^{-\hat{J}} R_{w\to v} + e^{\hat{J}}},$$
$$A_{w\to v} = \frac{e^{\hat{J}}(1 + m_{w\to v}) + e^{-\hat{J}}(1 - m_{w\to v})}{e^{-\hat{J}}(1 + m_{w\to v}) + e^{\hat{J}}(1 - m_{w\to v})}$$
$$= \frac{1 + \rho\, m_{w\to v}}{1 - \rho\, m_{w\to v}}. \tag{C.11}$$

Hence

$$R_{v\to p(v)} = e^{2u_v} \prod_{w\in \mathrm{ch}(v)} \frac{1+\rho\, m_{w\to v}}{1-\rho\, m_{w\to v}}. \tag{C.12}$$

Applying $m = (R-1)/(R+1)$ yields (C.10). Taking logarithms of (C.12) and using

$$\operatorname{artanh}(x) = \frac{1}{2}\log\frac{1+x}{1-x} \qquad \text{for } |x|<1$$

gives

$$\frac{1}{2}\log R_{v\to p(v)} = u_v + \sum_{w\in \mathrm{ch}(v)} \operatorname{artanh}\big(\rho\, m_{w\to v}\big).$$

Finally, $m = (R-1)/(R+1) = \tanh\left(\frac{1}{2}\log R\right)$, yielding (C.9). □

Remark C.5. Equation (C.9) is exactly the sum-product / belief-propagation update on the tree posterior (C.1). The point of the derivation above is that exactness follows directly from constrained subtree partition functions; no appeal to variational approximations is needed.

C.3 Root marginal

The root marginal is obtained by the same factorization argument, except that the root has no parent.

Proposition C.6 (Exact root magnetization)**.** *Let r be the root of T. Define*

$$Z_r^\sigma := \sum_{\substack{s\in\{-1,+1\}^V \\ s_r=\sigma}} \exp\left(\hat{J}\sum_{(u,v)\in E} s_u s_v + \sum_{v\in V} u_v s_v\right) \tag{C.13}$$

Then the posterior mean of the root spin is

$$\begin{aligned} H_r &:= u_r + \sum_{w\in \mathrm{ch}(r)} \operatorname{artanh}\big(\rho\, m_{w\to r}\big), \\ \langle s_r\rangle &= \frac{Z_r^+ - Z_r^-}{Z_r^+ + Z_r^-} = \tanh(H_r). \end{aligned} \tag{C.14}$$

Proof. The same factorization used in Proposition C.1 gives

$$Z_r^\sigma = e^{u_r\sigma} \prod_{w\in \mathrm{ch}(r)} \left(\sum_{\tau\in\{-1,+1\}} e^{\hat{J}\sigma\tau} Z_{w\to r}^\tau\right).$$

The ratio Z_r^+/Z_r^- therefore obeys the same algebra as in Proposition C.3, and the final formula follows from the same transformation as in Proposition C.4. □

C.4 Homogeneous regular-tree reduction

We now specialize to the model analyzed in Section 5: the regular b-ary tree $T_{b,d}$ with homogeneous local field

$$u_v \equiv u \qquad \text{for all } v\in V_{b,d}.$$

By symmetry, all cavity messages sent from subtrees of the same height are identical.

For $t\ge 0$, let T_t denote the rooted regular b-ary tree of depth t with homogeneous field u. Let m_t denote the posterior mean of the root spin of T_t:

$$m_t := \langle s_\varnothing\rangle \quad \text{under the Gibbs law on } T_t. \tag{C.15}$$

Because removing the parent edge of a depth-$t+1$ subtree produces a depth-t tree of the same form, the cavity message from a depth-t child subtree equals m_t. Therefore the general root recursion closes exactly on the one-dimensional sequence $(m_t)_{t\ge 0}$.

Proposition C.7 (Exact homogeneous recursion)**.** *For the regular b-ary tree with homogeneous field u,*

$$m_0 = \tanh(u), \tag{C.16}$$

and for every $t\ge 0$,

$$m_{t+1} = F_u(m_t) := \tanh\big(u + b\,\operatorname{artanh}\big(\rho\, m_t\big)\big). \tag{C.17}$$

Equivalently,

$$m_{t+1} = \frac{e^{2u}(1+\rho\, m_t)^b - (1-\rho\, m_t)^b}{e^{2u}(1+\rho\, m_t)^b + (1-\rho\, m_t)^b}. \tag{C.18}$$

Proof. For $t=0$, the tree consists of a single root with Gibbs weight proportional to $e^{us_\varnothing}$, so

$$m_0 = \frac{e^u - e^{-u}}{e^u + e^{-u}} = \tanh(u).$$

Now assume $t\ge 0$. A depth-$(t+1)$ tree consists of a root whose b children each support a depth-t subtree. By symmetry, every incoming child message equals m_t. Substituting these identical messages into Proposition C.6 gives

$$\begin{aligned} m_{t+1} &= \tanh\left(u + \sum_{i=1}^{b} \operatorname{artanh}\big(\rho\, m_t\big)\right) \\ &= \tanh\big(u + b\,\operatorname{artanh}\big(\rho\, m_t\big)\big), \end{aligned}$$

which is (C.17). Equation (C.18) is obtained by applying the identity

$$\tanh\left(\frac{1}{2}\log A\right) = \frac{A-1}{A+1}$$

to the ratio form inherited from (C.10). □

Remark C.8 (Ratio form). It is sometimes convenient to work with the depth-indexed partition-function ratios

$$R_t := \frac{Z_t^+}{Z_t^-}, \tag{C.19}$$

where $Z_t^\pm$ are the root-constrained partition functions of the homogeneous depth-t tree. Then

$$R_0 = e^{2u}, \qquad R_{t+1} = e^{2u}\left(\frac{e^{\hat{J}}R_t + e^{-\hat{J}}}{e^{-\hat{J}}R_t + e^{\hat{J}}}\right)^b, \tag{C.20}$$

and $m_t = (R_t - 1)/(R_t + 1)$. We use the magnetization form in the main text because it linearizes naturally around $u = 0$.

Proposition C.7 is the exact finite-depth recursion used throughout Section 5. The susceptibility theorem and the hallucinated-repair threshold law are both consequences of analyzing the map F_u in (C.17) near the literal fixed point.

D Proof of the Finite-Depth Susceptibility Law

This appendix proves Theorem 5.2. The result is a genuine linear-response theorem for the recursive repair grammar: the exact root magnetization on a depth-d tree has a first-order response coefficient equal to the geometric susceptibility

$$\chi_d = \sum_{\ell=0}^{d} (b\rho)^\ell, \qquad \rho = \tanh(\hat{J}),$$

and its first nonlinear correction is cubic in the homogeneous field u. The proof is exact and finite-depth: no thermodynamic limit, saddle-point approximation, or uncontrolled truncation is used.

D.1 Setup

Fix branching factor $b \geq 1$ and student coupling $\hat{J} \in \mathbb{R}$, and let

$$\rho := \tanh(\hat{J}), \qquad |\rho| < 1.$$

For each depth $t \geq 0$, let $m_t(u)$ denote the exact root magnetization of the homogeneous depth-t tree from Proposition C.7. Thus

$$m_0(u) = \tanh(u), \tag{D.1}$$

and for every $t \geq 0$,

$$\begin{aligned} F(u, m) &:= \tanh\big(u + b\ \mathrm{artanh}(\rho\, m)\big), \\ m_{t+1}(u) &= F\big(u, m_t(u)\big). \end{aligned} \tag{D.2}$$

Theorem 5.2 asserts that, in the small-field regime,

$$m_d(u) = \chi_d u + O(u^3), \qquad \chi_d = \sum_{\ell=0}^{d} (b\rho)^\ell. \tag{D.3}$$

D.2 Analyticity and oddness

We first prove that each finite-depth magnetization is an odd real-analytic function of the field near the origin. This is what forces the leading nonlinear term to be cubic rather than quadratic.

Lemma D.1 (Local analyticity and oddness)**.** *For every $t \geq 0$, there exists $r_t > 0$ such that m_t is real-analytic on $(-r_t, r_t)$. Moreover,*

$$m_t(-u) = -m_t(u) \qquad \textit{for all } u \in (-r_t, r_t). \tag{D.4}$$

Proof. We argue by induction on t.

For $t = 0$, $m_0(u) = \tanh(u)$, which is entire and odd.

Assume now that m_t is real-analytic and odd on some interval $(-r_t, r_t)$. Since $m_t(0) = 0$ and m_t is continuous, there exists $0 < r_{t+1} \leq r_t$ such that

$$|\rho\, m_t(u)| < 1 \qquad \text{for all } u \in (-r_{t+1}, r_{t+1}).$$

Because $\mathrm{artanh}(x)$ is real-analytic on $(-1, 1)$, the composition

$$u \mapsto \mathrm{artanh}\big(\rho\, m_t(u)\big)$$

is real-analytic on $(-r_{t+1}, r_{t+1})$. Since $\tanh$ is entire, the map

$$u \mapsto \tanh\big(u + b\ \mathrm{artanh}\big(\rho\, m_t(u)\big)\big)$$

is also real-analytic on that interval. By (D.2), this is exactly $m_{t+1}(u)$.

To prove oddness, note first that F is odd in the pair (u, m):

$$\begin{aligned} F(-u, -m) &= \tanh\big(-u + b\ \mathrm{artanh}(-\rho m)\big) \\ &= \tanh\big(-\big(u + b\ \mathrm{artanh}(\rho m)\big)\big) \\ &= -F(u, m). \end{aligned}$$

Since m_t is odd by the induction hypothesis,

$$\begin{aligned} m_{t+1}(-u) &= F(-u, m_t(-u)) \\ &= F(-u, -m_t(u)) \\ &= -F(u, m_t(u)) \\ &= -m_{t+1}(u). \end{aligned}$$

This completes the induction. □

Lemma D.1 implies that each m_t has a convergent odd Taylor expansion around $u = 0$:

$$m_t(u) = a_{t,1}u + a_{t,3}u^3 + a_{t,5}u^5 + \cdots \qquad \text{for } |u| < r_t. \tag{D.5}$$

The coefficient $a_{t,1}$ is the finite-depth susceptibility.

D.3 Derivative recursion and exact susceptibility

Definition D.2 (Finite-depth susceptibility)**.** For each $t \geq 0$, define

$$\chi_t := m_t'(0). \tag{D.6}$$

Lemma D.3 (Exact susceptibility recursion)**.** *The susceptibilities satisfy*

$$\chi_0 = 1, \qquad \chi_{t+1} = 1 + b\rho\,\chi_t \qquad \textit{for all } t \geq 0. \tag{D.7}$$

Consequently,

$$\chi_t = \sum_{\ell=0}^{t} (b\rho)^\ell. \tag{D.8}$$

Proof. The base case is immediate from $m_0(u) = \tanh(u)$:

$$\chi_0 = m_0'(0) = 1.$$

Now fix $t \geq 0$. Since m_t is real-analytic near 0, the chain rule applies to (D.2). Write

$$G(u, m) := u + b\ \mathrm{artanh}(\rho\, m),$$
$$F(u, m) = \tanh(G(u, m)).$$

Then

$$\partial_u F(u, m) = \mathrm{sech}^2(G(u, m)),$$

and

$$\partial_m F(u, m) = \mathrm{sech}^2(G(u, m)) \cdot b \cdot \frac{\rho}{1 - (\rho m)^2}.$$

Evaluating at $(u, m) = (0, 0)$ gives

$$\partial_u F(0, 0) = 1, \qquad \partial_m F(0, 0) = b\rho. \tag{D.9}$$

Differentiating (D.2) and setting $u = 0$, we obtain

$$m_{t+1}'(0) = \partial_u F(0, m_t(0)) + \partial_m F(0, m_t(0))\, m_t'(0).$$

Because m_t is odd, $m_t(0) = 0$, so (D.9) yields

$$\chi_{t+1} = 1 + b\rho\,\chi_t.$$

This proves (D.7). Solving the linear recursion with initial condition $\chi_0 = 1$ gives

$$\chi_t = \sum_{\ell=0}^{t} (b\rho)^\ell,$$

which is (D.8). □

Remark D.4 (Equivalent closed forms). If $b\rho \neq 1$, then

$$\chi_t = \frac{1 - (b\rho)^{t+1}}{1 - b\rho}. \tag{D.10}$$

On the critical line $b\rho = 1$,

$$\chi_t = t + 1. \tag{D.11}$$

In the ferromagnetic regime $\hat{J} \geq 0$, hence $\rho \geq 0$, this immediately yields the three scaling regimes emphasized in the main text: bounded response for $b\rho < 1$, linear growth for $b\rho = 1$, and exponential growth for $b\rho > 1$.

D.4 Cubic remainder bound

We now prove that the first nonlinear correction to the susceptibility expansion is cubic.

Lemma D.5 (Cubic remainder at fixed depth)**.** *For every $t \geq 0$, there exist constants $r_t^* > 0$ and $B_t < \infty$ such that*

$$|m_t(u) - \chi_t u| \leq B_t |u|^3 \qquad \textit{for all } |u| \leq r_t^*. \tag{D.12}$$

Proof. By Lemma D.1, m_t is real-analytic in a neighborhood of 0, and by oddness its Taylor series contains only odd powers:

$$m_t(u) = \chi_t u + a_{t,3}u^3 + a_{t,5}u^5 + \cdots \qquad \text{for } |u| < r_t.$$

Equivalently,

$$m_t(u) - \chi_t u = u^3 g_t(u),$$

where g_t is real-analytic near 0. In particular, there exists $r_t^* \in (0, r_t)$ such that g_t is bounded on $[-r_t^*, r_t^*]$. Setting

$$B_t := \sup_{|u| \leq r_t^*} |g_t(u)| < \infty$$

gives

$$\begin{aligned} |m_t(u) - \chi_t u| &= |u|^3 |g_t(u)| \\ &\leq B_t |u|^3, \\ &\text{for all } |u| \leq r_t^*. \end{aligned}$$

which proves (D.12). □

The preceding lemma is the precise reason the expansion in Theorem 5.2 has no quadratic term: oddness is not an aesthetic property but a structural consequence of the symmetric binary repair encoding.

D.5 Proof of Theorem 5.2

We now combine the derivative formula with the cubic remainder bound.

Proof of Theorem 5.2. Fix a depth $d \geq 0$. By Lemma D.3,

$$\chi_d = m_d'(0) = \sum_{\ell=0}^{d} (b\rho)^\ell.$$

By Lemma D.5 with $t = d$, there exist constants $r_d^* > 0$ and $B_d < \infty$ such that

$$\begin{aligned} m_d(u) &= \chi_d u + \varepsilon_d(u), \\ |\varepsilon_d(u)| &\leq B_d |u|^3, \\ &\quad \text{for } |u| \leq r_d^*. \end{aligned}$$

Hence

$$m_d(u) = \left(\sum_{\ell=0}^{d} (b\rho)^\ell \right) u + O(u^3) \qquad \text{as } u \to 0,$$

which is exactly the claimed finite-depth susceptibility law. □

Remark D.6 (Uniform finite-depth version). For a fixed maximal depth D, one may strengthen the result slightly by setting

$$u_*(D) := \min_{0 \leq t \leq D} r_t^*, \qquad B(D) := \max_{0 \leq t \leq D} B_t,$$

which gives the uniform bound

$$\begin{aligned} |m_t(u) - \chi_t u| &\leq B(D)|u|^3, \\ &\quad \text{for all } 0 \leq t \leq D,\ |u| \leq u_*(D). \end{aligned}$$

This is not needed for the statement of Theorem 5.2, but it is useful in the threshold analysis of Appendix E.

E Proof of the Hallucinated-Repair Threshold Law

This appendix proves Proposition 5.1 and Theorem 5.3. The key point is that the threshold law is *not* a statement about a purely homogeneous external field. Instead, it compares a *root-local literal bias* against a *distributed mismatch field* that acts across the recursive repair sites. This is the regime relevant to hallucinated repair: the clean ambiguous input locally favors the literal parse, while model mismatch globally tilts the recursive posterior toward repair.

Throughout this appendix we work in the ferromagnetic regime

$$\hat{J} \geq 0, \qquad \rho = \tanh(\hat{J}) \in [0, 1). \tag{E.1}$$

All statements are finite-depth and exact.

E.1 Mixed-field response function

For each depth $d \geq 0$, define the *mixed-field* root magnetization

$$M_d(u_0, \Delta)$$

as the posterior mean of the root spin on the regular tree $T_{b,d}$ when the root field is

$$u_\varnothing = u_0 + \Delta$$

and every non-root field is

$$u_v = \Delta, \qquad v \neq \varnothing.$$

Here $u_0 < 0$ is the teacher-consistent root-local literal bias of the clean ambiguous input, while $\Delta \geq 0$ is the global mismatch field induced by the student's miscalibrated prior/channel pair.

Let $m_t(\Delta)$ denote the homogeneous depth-t magnetization from Appendix D, i.e.,

$$\begin{aligned} m_t(\Delta) &= \langle s_\varnothing \rangle, \\ &\quad \text{for a depth-}t \text{ tree with } u_v \equiv \Delta \text{ for all } v. \end{aligned}$$

Lemma E.1 (Exact mixed-field reduction)**.** *For every $d \geq 1$,*

$$\begin{aligned} A_{d-1}(\Delta) &:= \operatorname{artanh}\big(\rho\, m_{d-1}(\Delta)\big), \\ M_d(u_0, \Delta) &= \tanh\big(u_0 + \Delta + b\, A_{d-1}(\Delta)\big). \end{aligned} \tag{E.2}$$

For $d = 0$,

$$M_0(u_0, \Delta) = \tanh(u_0 + \Delta). \tag{E.3}$$

Proof. For $d = 0$, the tree consists only of the root, so (E.3) is immediate.

Now let $d \geq 1$. By Proposition C.6, the exact root magnetization on any tree equals

$$\tanh\left(u_\varnothing + \sum_{w \in \mathrm{ch}(\varnothing)} \operatorname{artanh}\big(\rho\, m_{w \to \varnothing}\big) \right),$$

where $m_{w \to \varnothing}$ is the incoming cavity magnetization from child w. In the present mixed-field model, every child subtree is a depth-$(d-1)$ regular b-ary tree with homogeneous field Δ. By symmetry, each incoming message equals $m_{d-1}(\Delta)$. Substituting $u_\varnothing = u_0 + \Delta$ and summing over the b identical children yields (E.2). □

Define the *effective root driving field*

$$G_d(\Delta) := \Delta + b \operatorname{artanh}\big(\rho\, m_{d-1}(\Delta)\big), \qquad d \geq 1, \tag{E.4}$$

and for $d = 0$, set $G_0(\Delta) := \Delta$. Then

$$M_d(u_0, \Delta) = \tanh\big(u_0 + G_d(\Delta)\big). \tag{E.5}$$

Because $\tanh$ is strictly increasing, the sign of the Bayes decision is entirely determined by the sign of $u_0 + G_d(\Delta)$.

E.2 Monotonicity

We now prove the monotonicity facts used in the threshold argument.

Lemma E.2 (Ferromagnetic monotonicity of the homogeneous response)**.** *Assume* (E.1)*. For every* $t \geq 0$*, the map*

$$u \longmapsto m_t(u)$$

is odd and strictly increasing on its domain of definition.

Proof. Oddness was already established in Lemma D.1. It remains to prove strict monotonicity.

We argue by induction on t. For $t = 0$, $m_0(u) = \tanh(u)$, which is strictly increasing.

Assume m_t is strictly increasing. By Proposition C.7,

$$m_{t+1}(u) = \tanh\big(u + b \operatorname{artanh}\big(\rho\, m_t(u)\big)\big)\,.$$

Because $0 \leq \rho < 1$, the map

$$m \mapsto \operatorname{artanh}(\rho m)$$

is strictly increasing on $(-1, 1)$. Since $m_t(u) \in (-1, 1)$ and is strictly increasing by the induction hypothesis, the inner argument

$$u \mapsto u + b \operatorname{artanh}\big(\rho\, m_t(u)\big)$$

is strictly increasing. Composing with the strictly increasing $\tanh$ yields that m_{t+1} is strictly increasing as well. □

Lemma E.3 (Monotonicity of the mixed-field response)**.** *Assume* (E.1)*. For every* $d \geq 0$*:*

1. *the map* $\Delta \mapsto G_d(\Delta)$ *is odd and strictly increasing;*
2. *the mixed-field magnetization* $M_d(u_0, \Delta)$ *is strictly increasing in each argument* u_0 *and* Δ*.*

Proof. For $d = 0$, $G_0(\Delta) = \Delta$, so the claim is immediate.

Now let $d \geq 1$. Oddness of m_{d-1} from Lemma D.1, together with oddness of artanh, implies that G_d is odd. Strict monotonicity follows from Lemma E.2: if $\Delta_1 < \Delta_2$, then

$$m_{d-1}(\Delta_1) < m_{d-1}(\Delta_2),$$

hence

$$\operatorname{artanh}\big(\rho\, m_{d-1}(\Delta_1)\big) < \operatorname{artanh}\big(\rho\, m_{d-1}(\Delta_2)\big),$$

and therefore $G_d(\Delta_1) < G_d(\Delta_2)$.

Finally, by (E.5),

$$M_d(u_0, \Delta) = \tanh\big(u_0 + G_d(\Delta)\big).$$

Since $\tanh$ is strictly increasing, M_d is strictly increasing in u_0. Since G_d is strictly increasing in Δ, M_d is strictly increasing in Δ as well. □

Proof of Proposition 5.1. The homogeneous monotonicity statement follows from Lemma E.2. The mixed-field monotonicity statement follows from Lemma E.3. □

E.3 Cubic expansion of the effective driving field

We now identify the first nontrivial expansion of G_d near $\Delta = 0$.

Lemma E.4 (Small-mismatch expansion of G_d)**.** *Fix* $d \geq 0$*. There exist constants* $r_d > 0$ *and* $B_d < \infty$ *such that for all* $|\Delta| \leq r_d$*,*

$$G_d(\Delta) = \chi_d\, \Delta + R_d(\Delta), \qquad |R_d(\Delta)| \leq B_d |\Delta|^3, \tag{E.6}$$

where

$$\chi_d = \sum_{\ell=0}^{d} (b\rho)^\ell.$$

Proof. For $d = 0$, $G_0(\Delta) = \Delta$, so the claim holds with $\chi_0 = 1$ and $R_0 \equiv 0$.

Now let $d \geq 1$. By Theorem 5.2,

$$\begin{aligned} m_{d-1}(\Delta) &= \chi_{d-1}\, \Delta + \varepsilon_{d-1}(\Delta), \\ |\varepsilon_{d-1}(\Delta)| &\leq C_{d-1} |\Delta|^3. \end{aligned}$$

for $|\Delta|$ sufficiently small. Since $\rho \in [0, 1)$, there exists $r_d > 0$ such that

$$|\rho\, m_{d-1}(\Delta)| \leq \frac{1}{2} \qquad \text{for all } |\Delta| \leq r_d.$$

On $[-1/2, 1/2]$, the function $\operatorname{artanh}(x) - x$ is analytic and odd, hence bounded by a cubic term:

$$|\operatorname{artanh}(x) - x| \le C_{\text{atanh}}|x|^3.$$

Therefore

$$\operatorname{artanh}\big(\rho\, m_{d-1}(\Delta)\big) = \rho\, m_{d-1}(\Delta) + \delta_d(\Delta), \tag{E.7}$$

where

$$|\delta_d(\Delta)| \le C'_d|\Delta|^3 \qquad \text{for } |\Delta| \le r_d.$$

Substituting the expansion of m_{d-1} into (E.4) gives

$$\begin{aligned} G_d(\Delta) &= \Delta + b\rho\, m_{d-1}(\Delta) + b\,\delta_d(\Delta) \\ &= \Delta + b\rho\big(\chi_{d-1}\Delta + \varepsilon_{d-1}(\Delta)\big) + b\,\delta_d(\Delta) \\ &= \big(1 + b\rho\,\chi_{d-1}\big)\Delta + R_d(\Delta), \end{aligned}$$

where R_d absorbs the cubic remainders. By Lemma D.3,

$$1 + b\rho\,\chi_{d-1} = \chi_d,$$

so (E.6) follows. □

E.4 Existence, uniqueness, and asymptotics of the threshold

We now prove the main threshold theorem. Define the critical mismatch

$$\Delta_{\mathrm{c},d}(u_0) := \inf\{\Delta \ge 0 : M_d(u_0, \Delta) \ge 0\}, \qquad u_0 < 0. \tag{E.8}$$

Proof of Theorem 5.3. Fix $d \ge 0$ and $u_0 < 0$. Let

$$y := |u_0| > 0.$$

By (E.5),

$$\begin{aligned} M_d(u_0, \Delta) \ge 0 &\iff u_0 + G_d(\Delta) \ge 0 \\ &\iff G_d(\Delta) \ge y. \end{aligned}$$

Since G_d is continuous, odd, and strictly increasing by Lemma E.3, the threshold is characterized by the unique equation

$$G_d\big(\Delta_{\mathrm{c},d}(u_0)\big) = y \tag{E.9}$$

whenever a solution exists.

It remains to show that for small enough y, such a solution exists uniquely and obeys the claimed asymptotic law. By Lemma E.4, there exist $r_d > 0$ and $B_d < \infty$ such that

$$|G_d(\Delta) - \chi_d\Delta| \le B_d|\Delta|^3 \qquad \text{for all } |\Delta| \le r_d. \tag{E.10}$$

Choose

$$0 < y \le y_d^* := \min\left\{\frac{\chi_d r_d}{2}, \sqrt{\frac{\chi_d^3}{8B_d}}\right\}. \tag{E.11}$$

Set

$$\Delta_+ := \frac{2y}{\chi_d}.$$

Then $\Delta_+ \le r_d$, and using (E.10),

$$\begin{aligned} G_d(\Delta_+) &\ge \chi_d\Delta_+ - B_d\Delta_+^3 \\ &= 2y - B_d\left(\frac{2y}{\chi_d}\right)^3 \ge 2y - y = y. \end{aligned}$$

Since $G_d(0) = 0 < y$ and G_d is continuous and strictly increasing, there exists a unique

$$\Delta_{\mathrm{c},d}(u_0) \in [0, \Delta_+]$$

satisfying (E.9). This proves existence and uniqueness for sufficiently small $|u_0|$.

To obtain the asymptotic formula, write

$$G_d(\Delta) = \chi_d\Delta + R_d(\Delta)$$

with $|R_d(\Delta)| \le B_d|\Delta|^3$. Evaluating at $\Delta = \Delta_{\mathrm{c},d}(u_0)$ and using (E.9),

$$y = \chi_d\,\Delta_{\mathrm{c},d}(u_0) + R_d\big(\Delta_{\mathrm{c},d}(u_0)\big),$$

hence

$$\Delta_{\mathrm{c},d}(u_0) - \frac{y}{\chi_d} = -\frac{R_d\big(\Delta_{\mathrm{c},d}(u_0)\big)}{\chi_d}. \tag{E.12}$$

Because $\Delta_{\mathrm{c},d}(u_0) \le 2y/\chi_d$,

$$|R_d\big(\Delta_{\mathrm{c},d}(u_0)\big)| \le B_d\left(\frac{2y}{\chi_d}\right)^3.$$

Substituting into (E.12) gives

$$|\Delta_{\mathrm{c},d}(u_0) - \frac{y}{\chi_d}| \le \frac{8B_d y^3}{\chi_d^4} \le \frac{8B_d y^3}{\chi_d^3},$$

since $\chi_d \ge 1$ in the ferromagnetic regime. Recalling $y = |u_0|$, we conclude that

$$\Delta_{\mathrm{c},d}(u_0) = \frac{|u_0|}{\chi_d} + O\left(\frac{|u_0|^3}{\chi_d^3}\right) \qquad \text{as } u_0 \uparrow 0. \tag{E.13}$$

This is exactly the threshold law claimed in Theorem 5.3. □

Remark E.5 (Interpretation). Equation (E.13) is the formal statement that recursive structure amplifies mismatch. The clean ambiguous input contributes only a root-local literal bias u_0, whereas the mismatch field propagates through the full recursive tree and is amplified by the susceptibility χ_d. In the subcritical regime $b\rho < 1$, χ_d stays bounded, so an $O(1)$ mismatch is needed to induce false repair. On the critical line $b\rho = 1$, $\chi_d = d + 1$, so the threshold decays as $1/(d + 1)$. In the supercritical regime $b\rho > 1$, χ_d grows exponentially with depth, and the mismatch required to trigger hallucinated repair decays exponentially.

F Experimental Protocol for Exact Synthetic Validation

This appendix specifies the synthetic validation protocol used for the exact identities of Section 3 and the finite-depth laws of Section 5. The protocol is intentionally designed to isolate *teacher–student mismatch* from *algorithmic approximation*. Accordingly, all posterior quantities in Experiments F.2–F.3.2 are computed by exact dynamic programming on trees, rather than by approximate inference, which is the confound addressed in resource-rational noisy-channel models such as Clark et al. (2025). The only Monte Carlo in this appendix is over teacher-generated instances; inference itself is exact.

F.1 Exact inference backend and implementation checks

All synthetic computations are carried out on the finite recursive repair family of Section 4. For a tree $T = (V, E)$ with $|V| = n$, exact upward–downward sum-product inference on the posterior (4.3) has time complexity $O(n)$ and memory complexity $O(n)$, because the latent state alphabet is binary and the factor graph is a tree. Exact root and node marginals are computed from the cavity recursions in Appendix C; exact teacher sampling is performed by forward-filtering backward-sampling on the corresponding teacher tree model.

For the teacher prior

$$P_\theta(s \mid c) \propto \exp\left(J \sum_{(u,v)\in E} s_u s_v + h(c) \sum_{v\in V} s_v \right),$$

we first compute the teacher upward messages $Z^{\pm}_{v\to p(v)}$ exactly as in Appendix C, but with no observed cues. We then sample the root exactly from its marginal and sample each child conditional on its parent using

$$A^{(\tau)}_{u\to w} := \sum_{\xi\in\{-1,+1\}} e^{J\tau\xi} Z^{\xi}_{u\to w}.$$

$$K_w(\tau \mid \sigma, c) := e^{J\sigma\tau + h(c)\tau} \prod_{u\in \mathrm{ch}(w)} A^{(\tau)}_{u\to w}.$$

$$\mathbb{P}(s_w = \tau \mid s_{p(w)} = \sigma, c) \propto K_w(\tau \mid \sigma, c). \quad \text{(F.1)}$$

Conditioned on the sampled latent assignment $s^\star$, observed cues are then drawn independently from the teacher channel.

To verify the implementation of the exact dynamic program, we perform exhaustive brute-force checks on all tree settings with $|V| \le 15$. For each such setting, we enumerate all $2^{|V|}$ latent assignments, compute the exact partition function, node marginals, root magnetization, and overlap observables directly, and compare them to the dynamic-programming outputs. A configuration is declared valid only if the maximum absolute discrepancy across these quantities is at most 10^{-12}. Near the critical line $b\rho \in [0.98, 1.02]$, we additionally rerun a representative subset of configurations in arbitrary precision and require agreement to at least 10^{-10} in absolute error. No approximate posterior computation is used anywhere in Appendix F.

F.2 Synthetic family for Nishimori-identity validation

Teacher channel. For the identity experiments, we use a binary cue alphabet $\mathcal{Y} = \{0, 1\}$ with symmetric log-odds parameter $\lambda > 0$:

$$Z_\lambda := 2\cosh(\lambda/2),$$
$$q_\lambda(y \mid s, c^\circ) := \frac{\exp\left(\frac{\lambda}{2}(2y-1)s\right)}{Z_\lambda},$$
$$y \in \{0,1\}, \quad s \in \{-1,+1\}. \quad \text{(F.2)}$$

For this channel,

$$\frac{1}{2}\log\frac{q_\lambda(y \mid +1, c^\circ)}{q_\lambda(y \mid -1, c^\circ)} = \frac{\lambda}{2}(2y-1). \quad \text{(F.3)}$$

Thus λ directly controls local cue informativeness.

Teacher and student parameterization. The teacher parameters are (J, h, λ). In matched runs, the student uses

$$(\hat{J}, \hat{h}, \hat{\lambda}) = (J, h, \lambda).$$

For off-manifold gap-opening runs, we isolate pure field mismatch by setting

$$\hat{J} = J, \qquad \hat{\lambda} = \lambda, \qquad \hat{h} = h + \Delta,$$

so that the mismatch control parameter is exactly Δ. This choice ensures that any deviation from the Nishimori identities is attributable to model mismatch rather than to a change in local cue likelihood ratios.

Overlap and calibration observables. All identity experiments use the signed overlap

$$R(s, s') := \frac{1}{|V|} \sum_{v \in V} s_v s'_v \in [-1, 1]. \tag{F.4}$$

For a teacher draw $(s^\star, y)$ and exact student posterior $\Pi_{\hat{\theta},\hat{\eta}}(\cdot \mid y, c^\circ)$, let

$$m_v(y) := \langle s_v \rangle .$$

Then the exact truth-overlap and replica-overlap are

$$q_0(s^\star, y) := \langle R(s^\star, S) \rangle = \frac{1}{|V|} \sum_{v \in V} s^\star_v \, m_v(y), \tag{F.5}$$

$$q_{12}(y) := \left\langle R(S^{(1)}, S^{(2)}) \right\rangle = \frac{1}{|V|} \sum_{v \in V} m_v(y)^2, \tag{F.6}$$

where the second identity in (F.6) uses conditional independence of posterior replicas.

To validate Corollary 3.3, we use the root-spin observable

$$\phi(s) = s_\varnothing.$$

The exact posterior variance is

$$V_{\text{post}}(y) = \langle \phi(S)^2 \rangle - \langle \phi(S) \rangle^2 = 1 - m_\varnothing(y)^2, \tag{F.7}$$

and the exact per-instance Bayes loss is

$$L_{\text{Bayes}}(s^\star, y) = \left(s^\star_\varnothing - m_\varnothing(y) \right)^2. \tag{F.8}$$

Parameter grid. We evaluate the matched and off-manifold identities on the following grid:

$$b \in \{1, 2, 3\}, \tag{F.9}$$

$$\mathcal{D}_1 := \{0, 1, 2, 4, 8, 16, 32\},$$
$$\mathcal{D}_2 := \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\},$$
$$\mathcal{D}_3 := \{0, 1, 2, 3, 4, 5, 6, 7\},$$

$$d \in \begin{cases} \mathcal{D}_1, & b = 1, \\ \mathcal{D}_2, & b = 2, \\ \mathcal{D}_3, & b = 3, \end{cases} \tag{F.10}$$

$$b\rho \in \{0.6,\ 1.0,\ 1.4\}, \tag{F.11}$$
$$\text{whenever } \rho = \tanh(J) < 0.95,$$

$$h \in \{-0.2,\ 0,\ 0.2\}, \tag{F.12}$$

$$\lambda \in \{0.4,\ 0.8\}, \tag{F.13}$$

$$\Delta \in \{0,\ 10^{-3},\ 3 \cdot 10^{-3},\ 10^{-2},\ 3 \cdot 10^{-2},\ 10^{-1}\}. \tag{F.14}$$

The main paper displays representative slices of this grid; the appendix reports the full grid.

Matched-manifold protocol. For each grid point with $\Delta = 0$, we generate $N = 10^4$ exact teacher draws $(s^\star_i, y_i)$ using the forward-filtering backward-sampling procedure above. For each draw, we compute the exact posterior under the matched student model and record

$$\begin{aligned} q_{0,i} &:= q_0(s^\star_i, y_i), \\ q_{12,i} &:= q_{12}(y_i), \\ L_i &:= L_{\text{Bayes}}(s^\star_i, y_i), \\ V_i &:= V_{\text{post}}(y_i). \end{aligned}$$

The empirical estimates of the two exact identities are

$$\widehat{\Gamma}_R := \frac{1}{N} \sum_{i=1}^{N} (q_{12,i} - q_{0,i}), \tag{F.15}$$

$$\widehat{\Delta}_{\text{var}} := \frac{1}{N} \sum_{i=1}^{N} (L_i - V_i). \tag{F.16}$$

Theorems 3.1 and Corollaries 3.2–3.3 predict that both quantities are exactly zero in expectation. We report 95% percentile bootstrap confidence intervals based on 10^4 bootstrap resamples of the N teacher draws.

Off-manifold gap-opening protocol. For each $\Delta > 0$ in (F.14), we keep teacher generation fixed and evaluate the exact student posterior under $\hat{h} =$

$h + \Delta$. We then compute

$$\widehat{\Gamma}_R(\Delta) := \frac{1}{N}\sum_{i=1}^{N}\big(q_{12,i}^{(\Delta)} - q_{0,i}^{(\Delta)}\big), \tag{F.17}$$

where the superscript indicates that the posterior is evaluated under the mismatched student. The main paper uses these curves to show that the Nishimori gap is numerically zero on the matched manifold and opens immediately off it.

F.3 Synthetic protocol for susceptibility and threshold laws

Experiments F.3.1 and F.3.2 are deterministic: no teacher sampling is needed, because Theorem 5.2 and Theorem 5.3 are exact statements about the posterior response as a function of the effective field. We therefore work directly with the exact recursion of Appendix C.

F.3.1 Exact validation of the susceptibility law

For each (b, d, ρ) in the grid

$$b \in \{1, 2, 3\}, \tag{F.18}$$

$$d \in \begin{cases} \{0, 1, 2, 4, 8, 16, 32\}, & b = 1, \\ \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}, & b = 2, \\ \{0, 1, 2, 3, 4, 5, 6, 7\}, & b = 3, \end{cases} \tag{F.19}$$

$$b\rho \in \{0.6,\ 0.9,\ 1.0,\ 1.1,\ 1.4\}, \tag{F.20}$$

we compute the exact magnetization $m_d(u)$ at the symmetric small-field grid

$$u \in \mathcal{U} := \{\pm 2^{-k} : k = 4, 5, \ldots, 10\}. \tag{F.21}$$

We then fit the odd cubic model

$$m_d(u) \approx a_1 u + a_3 u^3 \tag{F.22}$$

by least squares over $u \in \mathcal{U}$, and define the empirical susceptibility

$$\widehat{\chi}_d := a_1. \tag{F.23}$$

The theorem prediction is

$$\chi_d = \sum_{\ell=0}^{d} (b\rho)^{\ell}. \tag{F.24}$$

To guard against numerical artifacts in the cubic fit, we perform two pre-registered consistency checks:

1. a central-difference estimate

$$\widehat{\chi}_d^{\mathrm{FD}} := \frac{m_d(\epsilon) - m_d(-\epsilon)}{2\epsilon}, \qquad \epsilon = 10^{-6}, \tag{F.25}$$

   must agree with $\widehat{\chi}_d$ to within relative error 10^{-6};

2. the cubic residual

$$\max_{u \in \mathcal{U}} \frac{|m_d(u) - \widehat{\chi}_d u|}{|u|^3} \tag{F.26}$$

   must remain bounded as $u \to 0$, consistent with the $O(u^3)$ remainder of Theorem 5.2.

The main quantitative outputs are:

- the absolute susceptibility error

$$E_{\mathrm{susc}}(b, d, \rho) := |\widehat{\chi}_d - \chi_d|; \tag{F.27}$$

- the regime-wise scaling plots of $\widehat{\chi}_d$ versus d in the subcritical, critical, and supercritical regimes.

F.3.2 Exact validation of the threshold law

To validate Theorem 5.3, we use the mixed-field response $M_d(u_0, \Delta)$ of Section 5 and Appendix E. For each (b, d, ρ) in the grid (F.18)–(F.20), and for each clean-bias magnitude

$$|u_0| \in \{2^{-3}, 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}, 2^{-8}, 2^{-9}, 2^{-10}\}, \tag{F.28}$$

we compute the unique exact threshold

$$\widehat{\Delta_{\mathrm{c}}}(d, u_0) = \inf\{\Delta \geq 0 : M_d(u_0, \Delta) \geq 0\} \tag{F.29}$$

by bisection. The bisection interval is initialized at

$$[0,\ 2|u_0|/\chi_d]$$

and expanded by repeated doubling if necessary, although the asymptotic theorem predicts that this initial interval is sufficient in the small-bias regime. The root-finding tolerance is

$$|M_d(u_0, \Delta)| \leq 10^{-12} \quad \text{or} \quad \text{interval width} \leq 10^{-12}.$$

The theorem predicts

$$\widehat{\Delta_{\mathrm{c}}}(d, u_0) = \frac{|u_0|}{\chi_d} + O\left(\frac{|u_0|^3}{\chi_d^3}\right). \tag{F.30}$$

We therefore report the following pre-specified diagnostics:

1. the collapse ratio

$$C(b, d, \rho, u_0) := \frac{\widehat{\Delta_c}(d, u_0)\, \chi_d}{|u_0|}, \qquad \text{(F.31)}$$

which should converge to 1 as $|u_0| \to 0$;

2. the normalized remainder

$$R(b, d, \rho, u_0) := \frac{\chi_d^3}{|u_0|^3} \left| \widehat{\Delta_c}(d, u_0) - \frac{|u_0|}{\chi_d} \right|, \qquad \text{(F.32)}$$

which should remain $O(1)$ in the small-bias regime;

3. in the supercritical regime $b\rho > 1$, the slope of

$$\log \widehat{\Delta_c}(d, u_0) \quad \text{versus} \quad d, \qquad \text{(F.33)}$$

which should approach $-\log(b\rho)$ at fixed u_0.

These diagnostics directly test the content of Theorem 5.3, rather than merely showing that the threshold is monotone.

F.4 Numerical precision, uncertainty quantification, and reporting rules

The synthetic protocol uses the following reporting rules.

Exactness checks. Every figure in the main paper and appendix is generated only from runs that pass the dynamic-programming versus brute-force validation threshold described in Section F.1. For critical-regime runs near $b\rho = 1$, representative slices are cross-checked in high precision.

Uncertainty quantification. Only the identity experiments involve Monte Carlo sampling over teacher draws, so only those experiments carry confidence intervals. All confidence intervals in the main paper are 95% percentile bootstrap intervals with 10^4 bootstrap resamples. Susceptibility and threshold experiments are deterministic given the parameter tuple and therefore do not receive bootstrap intervals; their numerical uncertainty is instead controlled by the implementation tolerances above.

No hidden model selection. No parameter settings are discarded after seeing the outputs. The grids in (F.9)–(F.14) and (F.18)–(F.28) are fixed before running the experiments. The main paper may visualize representative slices of these grids for space reasons, but the appendix reports full-grid summaries.

Reproducibility. All random seeds for teacher sampling and bootstrap resampling are fixed in advance and stored with the experiment configuration files. The released code will include the exact dynamic-programming routines, brute-force verification routines for small trees, and scripts generating every reported figure and table from the fixed parameter grids above.

G Natural-Language Symbolic Benchmark Construction

This appendix specifies the symbolic benchmark used in Section 6.3. The benchmark is not intended as a substitute for real human data. Its role is narrower and more controlled: it provides a bridge between the exact teacher–student theory of Sections 2–5 and human-readable sentence-like inputs. We focus on omission-sensitive function-word constructions and supportive-context manipulations because these are canonical pressure points for noisy-channel sentence processing, uncertain-input inference, and context-sensitive repair (Levy, 2008, 2011; Gibson et al., 2013; Futrell and Levy, 2017; Chen et al., 2023).

G.1 Combinatorial design of the benchmark

The benchmark is indexed by five factors:

$$\mathcal{B} = \mathcal{F} \times \mathcal{D} \times \mathcal{C} \times \mathcal{P} \times \mathcal{V},$$

where

$$\mathcal{F} = \{\textsc{comp}, \textsc{rel}, \textsc{inf}\},$$
$$\mathcal{D} = \{1, 2, 3, 4, 5, 6\}.$$

$$\mathcal{C} = \{\textsc{neutral}, \textsc{supportive}\}.$$

$$\mathcal{P} = \left\{ \begin{array}{l} \textsc{all-present}, \\ \textsc{root-omitted}, \\ \textsc{deep-omitted}, \\ \textsc{alternating}, \\ \textsc{edge-omitted}, \\ \textsc{all-omitted} \end{array} \right\}.$$

and

$$\mathcal{V} = \{0, 1, 2, 3, 4, 5\}.$$

Thus the reported benchmark contains

$$|\mathcal{B}| = 3 \times 6 \times 2 \times 6 \times 6 = 1296 \qquad \text{(G.1)}$$

distinct items.

The three construction families are:

1. **Complementizer omission** (COMP): nested complement clauses with optional complementizer *that*;

2. **Relative-pronoun omission** (REL): nested relative-clause templates with optional relative pronoun *who*;

3. **Infinitival-marker omission** (INF): nested infinitival templates with optional infinitival marker *to*.

Each family contains a NEUTRAL preamble and a SUPPORTIVE preamble. The supportive preamble is designed to tilt the student toward expecting omissions, in the same spirit as supportive discourse context or supportive corruption ecology in noisy-channel sentence processing (Chen et al., 2023; Ryskin et al., 2018). The benchmark therefore combines local marker evidence, recursive depth, and contextually modulated omission expectations in a fully controlled symbolic setting.

G.2 Template realization and omission patterns

For an item of depth d, let

$$r = (r_1, \ldots, r_d) \in \{0, 1\}^d$$

be the marker-presence vector, where $r_i = 1$ means that the family-specific function word is overtly present at recursion level i, and $r_i = 0$ means it is omitted. The six pattern classes are defined deterministically as follows:

$$\begin{aligned}
\text{ALL-PRESENT}: &\quad r_i = 1 \quad \forall i, && \text{(G.2)}\\
\text{ROOT-OMITTED}: &\quad r_1 = 0, && \text{(G.3)}\\
&\quad r_i = 1 \quad \text{for } i > 1,\\
\text{DEEP-OMITTED}: &\quad r_d = 0, && \text{(G.4)}\\
&\quad r_i = 1 \quad \text{for } i < d,\\
\text{ALTERNATING}: &\quad r_i = \mathbb{I}\{\{\}\, i \text{ is even}\}, && \text{(G.5)}\\
\text{EDGE-OMITTED}: &\quad r_1 = 0 \quad \text{if } d = 1, && \text{(G.6)}\\
&\quad r_1 = r_d = 0 \quad \text{if } d > 1,\\
&\quad r_i = 1 \quad \text{for } 1 < i < d \text{ and } d > 1\\
\text{ALL-OMITTED}: &\quad r_i = 0 \quad \forall i. && \text{(G.7)}
\end{aligned}$$

The observed sentence $x^{\text{obs}}_{f,d,c,r,v}$ is produced by instantiating a family-specific recursive template and inserting the family marker only at positions with $r_i = 1$. The intended sentence $x^{\star}_{f,d,c,v}$ uses the same lexical content but with all family markers present. Thus every benchmark item comes with an *observed* surface form and an *intended repaired* counterpart. Concretely, the three families instantiate the following schematic forms:

$$\begin{aligned}
\text{COMP}: &\quad \text{Prefix}(c)\ \text{NP}_1\text{V}_1\ [\texttt{that}] && \text{(G.8)}\\
&\quad \text{NP}_2\text{V}_2\ [\texttt{that}]\ \cdots\ \text{Clause}\\
\text{REL}: &\quad \text{Prefix}(c)\ \text{NP}_1\ [\texttt{who}] && \text{(G.9)}\\
&\quad \text{NP}_2\ [\texttt{who}]\ \cdots\ \text{VP}\\
\text{INF}: &\quad \text{Prefix}(c)\ \text{NP}\ \text{V}_1\ [\texttt{to}] && \text{(G.10)}\\
&\quad \text{V}_2\ [\texttt{to}]\ \cdots\ \text{V}_{d+1}
\end{aligned}$$

Brackets indicate optional overt markers controlled by r. Family-specific lexicalizations are drawn from finite lexicons of subjects, verbs, clause termini, and head nouns, yielding six lexical variants per factor configuration.

G.3 Teacher model for symbolic items

The symbolic benchmark specializes the recursive repair grammar to a chain ($b = 1$), which is the natural latent structure for nested clause templates. For a fixed item $b = (f, d, c, r, v)$, let

$$s = (s_1, \ldots, s_d) \in \{-1, +1\}^d$$

be the latent binary repair assignment along the recursive chain, with s_1 denoting the root repair decision. The teacher distribution over s is

$$\begin{aligned}
H^{\text{T}}_f(s; b) &:= \sum_{i=1}^{d} u^{\text{T}}_{f,i}(b)\, s_i + \sum_{i=1}^{d-1} J^{\text{T}}_{f,i}(b)\, s_i s_{i+1},\\
P^{\text{T}}_f(s \mid b) &\propto \exp\!\big(H^{\text{T}}_f(s; b)\big).
\end{aligned} \tag{G.11}$$

The item-specific teacher fields are

$$\begin{aligned}
u^{\text{T}}_{f,i}(b) = \beta^{\text{ctx}}_f \mathbb{I}\{c = \text{SUPPORTIVE}\} + \beta^{\text{cue}}_f(r_i)\, \tau_f(i, d)\\
+ \lambda_f\, \delta_{v,i} + \beta^{\text{root}}_f \mathbb{I}\{i = 1\}\,.
\end{aligned} \tag{G.12}$$

where

$$\beta^{\text{cue}}_f(r_i) = \begin{cases} \beta^{\text{pres}}_f, & r_i = 1,\\ \beta^{\text{omit}}_f, & r_i = 0, \end{cases} \tag{G.13}$$

and the depth taper is

$$\tau_f(i, d) = \begin{cases} 1, & d = 1,\\ 1 - \rho_f \dfrac{i-1}{d-1}, & d > 1. \end{cases} \tag{G.14}$$

The lexical offsets are taken from the fixed set

$$\Delta_{\text{lex}} = \left\{ \begin{matrix} -0.18,\ -0.11,\ -0.05,\ 0.00, \\ 0.05,\ 0.11,\ 0.18 \end{matrix} \right\}.$$

and are injected deterministically via

$$\delta_{v,i} = \Delta_{\text{lex}}[(v + 2(i-1)) \bmod 7]\,. \tag{G.15}$$

The teacher couplings are

$$J^{\text{T}}_{f,i}(b) = J^{\text{base}}_f + \gamma_f \frac{d-1-i}{\max(1, d-2)} + 0.02\,\delta'_{v,i}, \tag{G.16}$$

with

$$\delta'_{v,i} = \Delta_{\text{lex}}[(v + i - 1) \bmod 7]\,. \tag{G.17}$$

The last term is a small deterministic lexical perturbation included to avoid a perfectly degenerate benchmark.

Table 2 lists the family-specific teacher parameters used in the reported run.

G.4 Student models and exact inference

We compare three models on the same symbolic benchmark.

Recursive coupled noisy channel. The coupled student uses the same chain factorization as (G.11), but with family-specific student parameters:

$$P^{\text{C}}_f(s \mid b) \propto \exp\left(\sum_{i=1}^{d} u^{\text{C}}_{f,i}(b)\, s_i + \tilde{J}_f \sum_{i=1}^{d-1} s_i s_{i+1}\right), \tag{G.18}$$

where

$$u^{\text{C}}_{f,i}(b) = \tilde{\beta}^{\text{ctx}}_f \mathbb{I}\{c = \textsc{supportive}\} + \tilde{\beta}^{\text{cue}}_f(r_i) + \tilde{\beta}^{\text{root}}_f \mathbb{I}\{i = 1\}\,, \tag{G.19}$$

and

$$\tilde{\beta}^{\text{cue}}_f(r_i) = \begin{cases} \tilde{\beta}^{\text{pres}}_f, & r_i = 1, \\ \tilde{\beta}^{\text{omit}}_f, & r_i = 0. \end{cases} \tag{G.20}$$

Uncoupled noisy-channel baseline. The uncoupled baseline uses the same fields $u^{\text{C}}_{f,i}(b)$ but sets all couplings to zero:

$$P^{\text{U}}_f(s \mid b) \propto \exp\left(\sum_{i=1}^{d} u^{\text{C}}_{f,i}(b)\, s_i\right). \tag{G.21}$$

This baseline preserves the same local cue parameterization while deleting the recursive consistency term.

Literal-only baseline. The literal baseline assigns a fixed repair probability of 10^{-6} to every item:

$$p^{\text{L}}_{\text{repair}}(b) = 10^{-6}. \tag{G.22}$$

It therefore functions as a hard no-repair baseline.

Table 3 lists the coupled/uncoupled student parameters used in the reported run.

For all three models, the quantity of interest is the root repair probability. Because the latent structure is a chain, this probability is computed exactly by a one-dimensional backward recursion. Writing u_i for the item-specific fields and J_i for the chain couplings, define

$$m_d = \tanh(u_d), \tag{G.23}$$

and for $i = d-1, d-2, \ldots, 1$,

$$m_i = \tanh\big(u_i + \operatorname{artanh}\big(\tanh(J_i)\, m_{i+1}\big)\big)\,. \tag{G.24}$$

Then the exact root repair probability is

$$p_{\text{repair}}(b) = \frac{1 + m_1}{2}. \tag{G.25}$$

For the coupled student, $J_i = \tilde{J}_f$ for all i; for the uncoupled baseline, $J_i = 0$ for all i.

G.5 Teacher-sampled pseudo-observer responses

Section 6.3 uses teacher-sampled pseudo-observer responses rather than real participant data. For each item $b \in \mathcal{B}$, let

$$p^{\text{T}}_b$$

be the exact teacher root repair probability obtained from (G.25) using the teacher parameters. We then draw

$$K_b \sim \text{Binomial}(n_{\text{resp}}, p^{\text{T}}_b), \qquad n_{\text{resp}} = 40, \tag{G.26}$$

and define the teacher-sampled repair rate

$$\hat{r}_b := \frac{K_b}{40}. \tag{G.27}$$

This makes the benchmark stochastic at the response level while keeping the underlying teacher probabilities exact. The resulting pseudo-observer responses are therefore interpretable as a controlled finite-sample version of the teacher distribution, not as human annotations.

Table 2: Teacher parameters for the symbolic benchmark. For each family f, β_f^{root} is the root-local literal bias, β_f^{ctx} is the supportive-context shift, β_f^{omit} and β_f^{pres} are the omission and presence cue fields, ρ_f is the depth taper, λ_f is the lexical-scale coefficient, and $J_f^{\text{base}}, \gamma_f$ define the teacher coupling schedule in (G.16).

Family	β_f^{root}	β_f^{ctx}	β_f^{omit}	β_f^{pres}	ρ_f	λ_f	$(J_f^{\text{base}}, \gamma_f)$
Complementizer omission	-0.52	0.12	0.84	-0.62	0.10	0.08	$(1.02, 0.16)$
Relative-pronoun omission	-0.60	0.10	0.96	-0.70	0.12	0.07	$(0.92, 0.13)$
Infinitival-marker omission	-0.42	0.08	0.70	-0.46	0.08	0.06	$(0.74, 0.10)$

Table 3: Student parameters for the symbolic benchmark. The recursive coupled model uses $\tilde{J}_f$ as a constant chain coupling; the uncoupled baseline uses the same local fields but sets $\tilde{J}_f = 0$.

Family	$\tilde{\beta}_f^{\text{root}}$	$\tilde{\beta}_f^{\text{ctx}}$	$\tilde{\beta}_f^{\text{omit}}$	$\tilde{\beta}_f^{\text{pres}}$	$\tilde{J}_f$
Complementizer omission	-0.48	0.10	0.78	-0.56	0.96
Relative-pronoun omission	-0.56	0.09	0.88	-0.64	0.86
Infinitival-marker omission	-0.38	0.07	0.64	-0.42	0.70

G.6 Recursive-drive statistic and evaluation metrics

To connect the symbolic benchmark to the susceptibility theory, we define the chain susceptibility of family f at depth d by

$$\chi_d^{(f)} = \sum_{\ell=0}^{d-1} \tanh(\tilde{J}_f)^{\ell}. \tag{G.28}$$

Let

$$\omega_b = 1 - \frac{1}{d} \sum_{i=1}^{d} r_i \tag{G.29}$$

be the omitted-marker fraction of item b. The benchmark-level recursive-drive statistic is then

$$\Xi_b = \chi_d^{(f)} \, \omega_b \left(1 + \frac{1}{2} \mathbb{I}\{c = \text{SUPPORTIVE}\}\right). \tag{G.30}$$

This is the quantity used in Section 6.3 and Figure 3 to stratify items by effective recursive drive.

Given a model $M \in \{\text{L}, \text{U}, \text{C}\}$ with predicted repair probability $p_b^{(M)}$, we report the following benchmark metrics:

Majority-label accuracy. Let

$$y_b^{\text{maj}} = \mathbb{I}\left\{\hat{r}_b \geq \tfrac{1}{2}\right\}.$$

Then

$$\text{Acc}(M) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbb{I}\left\{\mathbb{I}\left\{p_b^{(M)} \geq \tfrac{1}{2}\right\} = y_b^{\text{maj}}\right\}. \tag{G.31}$$

Item-wise log loss.

$$\text{NLL}(M) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \Big[-\hat{r}_b \log p_b^{(M)} - (1 - \hat{r}_b) \log(1 - p_b^{(M)})\Big]. \tag{G.32}$$

Brier score.

$$\text{Brier}(M) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left(p_b^{(M)} - \hat{r}_b\right)^2. \tag{G.33}$$

Expected calibration error. Using 10 equal-width bins B_k over $[0, 1]$,

$$\text{ECE}(M) = \sum_{k=1}^{10} \frac{|B_k|}{|\mathcal{B}|} \left| \frac{1}{|B_k|} \sum_{b \in B_k} p_b^{(M)} - \frac{1}{|B_k|} \sum_{b \in B_k} \hat{r}_b \right|. \tag{G.34}$$

High-Ξ subset loss. Let

$$q_{0.75} = 75\text{th percentile of } \{\Xi_b : b \in \mathcal{B}\}.$$

Then the top-Ξ quartile score is

$$\mathcal{B}_{0.75} := \{b : \Xi_b \geq q_{0.75}\},$$

$$\text{NLL}_{\Xi \geq q_{0.75}}(M) = \frac{1}{|\mathcal{B}_{0.75}|} \sum_{b \in \mathcal{B}_{0.75}} \Big[-\hat{r}_b \log p_b^{(M)} - (1 - \hat{r}_b) \log(1 - p_b^{(M)})\Big]. \tag{G.35}$$

G.7 Reported configuration and relation to Figure 3

The reported run in Section 6.3 uses:

$$\begin{aligned} d_{\max} &= 6, \\ |\mathcal{V}| &= 6, \\ n_{\text{resp}} &= 40, \\ &\quad 8 \text{ calibration bins}, \\ &\quad 6\ \Xi\text{-bins for gain plots.} \end{aligned}$$

No parameter fitting is performed in Section 6.3; all 1296 items are evaluated exhaustively. Training/held-out splits are introduced only in the later adversarial-control and micro-study appendices, not in the symbolic sanity-check itself.

The three panels of Figure 3 correspond exactly to:

1. a representative depth curve for the COMP family with supportive context and the ALL-OMITTED pattern;
2. a global calibration plot comparing uncoupled and coupled models on the full benchmark;
3. a binned plot of coupled-model log-loss gain,

$$\text{gain}(b) = \text{NLL}_b(\text{U}) - \text{NLL}_b(\text{C}),$$

stratified by Ξ_b.

The key role of this appendix is therefore twofold. First, it makes the symbolic benchmark fully reproducible. Second, it clarifies that Section 6.3 is not a loose qualitative illustration: every symbolic item, every pseudo-observer response, every model probability, and every reported metric is generated by an explicit, exact, and finite protocol.

H Optional Extensions and Additional Robustness

This appendix collects three optional but high-leverage robustness experiments. Their purpose is not to rescue the main claim of the paper—Sections 3–5 and Section 6 already establish the exact theory and its direct numerical consequences—but to neutralize the most predictable reviewer objections. Concretely, these experiments address three distinct failure modes:

1. the objection that the coupled model wins only because it has effectively stronger local evidence rather than genuinely recursive structure;
2. the objection that the threshold law is an artifact of regular trees or a single stylized corruption channel;
3. the objection that the depth-sensitive predictions remain confined to teacher-generated synthetic data.

All three experiments use exact posterior inference within the relevant model class. Thus, as in the main text, the issue under test is model structure and mismatch, not approximate inference (Levy, 2008, 2011; Clark et al., 2025).

H.1 Full numerical summaries for the exact validations

H.2 Symbolic benchmark plots

H.3 Iso-local-field adversarial control

The first robustness experiment is designed to answer the strongest empirical objection to Section 6.3: perhaps the recursive coupled model improves only because it has access to stronger effective local evidence. To test this directly, we construct two adversarial local-only baselines and force them to compete against the recursive coupled model on held-out lexical variants.

Let $\mathcal{B}$ denote the symbolic benchmark from Appendix G. We split lexical variants into a training subset $\mathcal{V}_{\text{tr}}$ and a held-out subset $\mathcal{V}_{\text{te}}$. All evaluation metrics reported below are computed only on items with lexical variants in $\mathcal{V}_{\text{te}}$.

Iso-local-field uncoupled control. This control uses exactly the same local fields as the recursive coupled student from Appendix G.4, but sets all couplings to zero:

$$P_f^{\text{iso}}(s \mid b) \propto \exp\left(\sum_{i=1}^{d} u_{f,i}^{\text{C}}(b)\, s_i\right). \tag{H.1}$$

Because the local fields are matched, any gain of the coupled model over this baseline cannot be attributed to stronger pointwise cue evidence.

Shallow-calibrated local-only adversary. We further construct a deliberately strengthened local-only adversary by fitting additive field offsets only on shallow training items. For each family f, pattern class p, and context c, the adversary adds a scalar shift $\delta_{f,p,c}$ to every local field:

$$P_f^{\text{adv}}(s \mid b) \propto \exp\left(\sum_{i=1}^{d} \left(u_{f,i}^{\text{C}}(b) + \delta_{f,p,c}\right) s_i\right). \tag{H.2}$$

Table 4: Representative exact synthetic results for Section 6.1 on depth-8 trees with 511 nodes ($b = 2$), $h = 0$, and $\lambda = 0.8$. Bracketed values are 95% bootstrap confidence intervals. The matched-manifold residuals are statistically indistinguishable from 0, while the off-manifold Nishimori gap opens monotonically and most strongly in the supercritical regime.

Regime	$\alpha = b \tanh J$	$\hat{\Gamma}_R(0)$	$\widehat{\Delta}_{\mathrm{cal}}(0)$	$\hat{\Gamma}_R(0.03)$	$\hat{\Gamma}_R(0.10)$
Subcritical	0.60	0.0000 [-0.0004, 0.0004]	0.0003 [-0.0152, 0.0163]	0.0019 [0.0015, 0.0022]	0.0193 [0.0190, 0.0197]
Critical	1.00	-0.0003 [-0.0008, 0.0001]	-0.0062 [-0.0215, 0.0098]	0.0041 [0.0036, 0.0045]	0.0469 [0.0464, 0.0475]
Supercritical	1.40	-0.0005 [-0.0011, 0.0002]	-0.0003 [-0.0169, 0.0164]	0.0106 [0.0100, 0.0112]	0.1105 [0.1095, 0.1115]

Table 5: Representative exact synthetic results for Section 6.2 on depth-8 trees with $b = 2$ and $|u_0| = 2^{-8}$. Susceptibilities are estimated independently from an adaptive odd cubic fit $m_d(u) \approx a_1 u + a_3 u^3$ in the asymptotic small-field regime, and thresholds are computed independently by exact monotone bisection on the mixed-field response. The threshold theorem predicts $\Delta_c \approx |u_0|/\chi_d$, and both susceptibility and threshold relative errors are at numerical precision across all three regimes.

Regime	$\alpha = b \tanh J$	χ_8	$\hat{\chi}_8$	rel.err.(χ)	$\hat{\Delta}_c$	$2^{-8}/\chi_8$	rel.err.(Δ)
Subcritical	0.60	2.474806	2.474806	1.62×10^{-9}	0.001578	0.001578	6.07×10^{-6}
Critical	1.00	9.000000	9.000000	9.70×10^{-10}	0.000434	0.000434	6.78×10^{-6}
Supercritical	1.40	49.152617	49.152617	1.14×10^{-9}	0.000079	0.000079	2.25×10^{-6}

The offsets are fit by minimizing item-wise negative log likelihood on the shallow training subset

$$\mathcal{B}^{\mathrm{tr}}_{\leq 2} = \{b \in \mathcal{B} : d(b) \leq 2,\ v(b) \in \mathcal{V}_{\mathrm{tr}}\}.$$

Formally,

$$\begin{aligned} \mathcal{B}^{\mathrm{tr}}_2 &:= \mathcal{B}^{\mathrm{tr}}_{\leq 2}, \\ J_{\mathrm{adv}}(\delta) &:= \sum_{b \in \mathcal{B}^{\mathrm{tr}}_2} \Big[-\hat{r}_b \log p^{\mathrm{adv}}_b(\delta) \\ &\qquad - (1 - \hat{r}_b) \log\big(1 - p^{\mathrm{adv}}_b(\delta)\big) \Big], \\ \delta^\star &= \mathrm{argmin}_\delta\, J_{\mathrm{adv}}(\delta). \end{aligned} \tag{H.3}$$

This is intentionally adversarial: the local-only control is calibrated on precisely the shallow regime where a recursive model should be least necessary.

For each held-out item b, define the item-wise held-out gains

$$\begin{aligned} g^{\mathrm{iso}}_b &= \mathrm{NLL}_b(\mathrm{iso}) - \mathrm{NLL}_b(\mathrm{cpl}), \\ g^{\mathrm{adv}}_b &= \mathrm{NLL}_b(\mathrm{adv}) - \mathrm{NLL}_b(\mathrm{cpl}). \end{aligned} \tag{H.4}$$

where $\mathrm{NLL}_b(M)$ is the per-item contribution of model M to (G.32). If the recursive model is genuinely exploiting recursive structure rather than merely stronger local evidence, both gains should remain positive and should increase in the high-drive regime.

Figure 4 and Table 6 show that both adversarial baselines remain decisively worse than the recursive coupled model on held-out items. The shallow-calibrated local-only adversary does improve over the iso-local uncoupled control on shallow inputs, exactly as intended, but still fails on deep or high-Ξ items. This is the critical result: even after explicitly strengthening local-only evidence on the regime most favorable to it, the recursive coupled model remains best where the theory predicts recursive amplification should matter most.

H.4 Irregular-tree and channel-family robustness

The second robustness experiment tests whether the threshold law of Theorem 5.3 survives when we move beyond perfectly regular trees and beyond the binary-logit channel used in the main theorem family. The central object is the generalized susceptibility of an arbitrary finite rooted tree T:

$$\chi(T) := \left. \frac{\partial M_T(u)}{\partial u} \right|_{u=0}, \tag{H.5}$$

where $M_T(u)$ is the exact root magnetization under a homogeneous infinitesimal field u. In the experiments below, $\chi(T)$ is computed independently from exact linear-response recursions, not recovered from the threshold itself.

We consider three tree families:

1. regular rooted trees;
2. Galton–Watson trees truncated at finite depth;
3. template-derived irregular trees extracted from the symbolic benchmark topology.

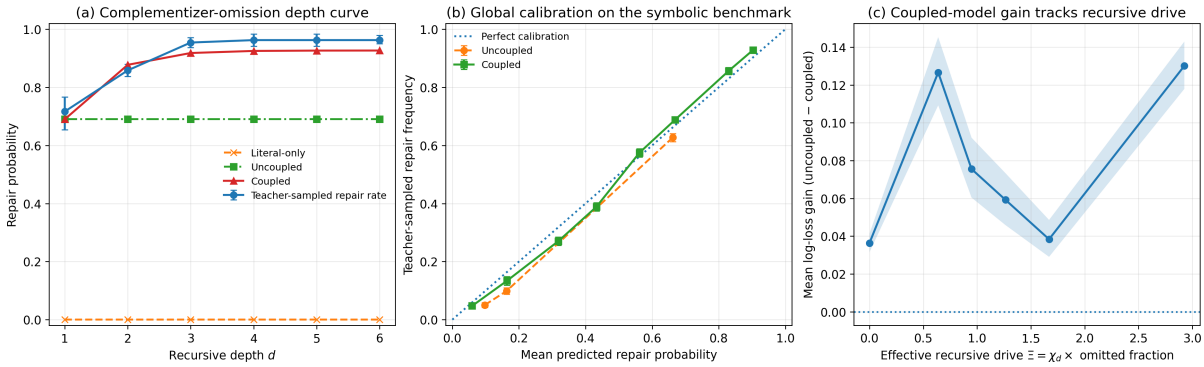

Figure 3: **Symbolic natural-language sanity check.** Left: on complementizer-omission templates, the recursive coupled model tracks teacher-sampled repair rates as recursive depth increases, whereas the literal-only and uncoupled baselines fail in systematically different ways. Middle: on the full symbolic benchmark, the coupled model is better calibrated to teacher-sampled repair frequencies than the uncoupled baseline. Right: coupled-model gain is concentrated at high effective recursive drive $\Xi = \chi_d \times$ omitted fraction, linking the symbolic benchmark back to the susceptibility theory.

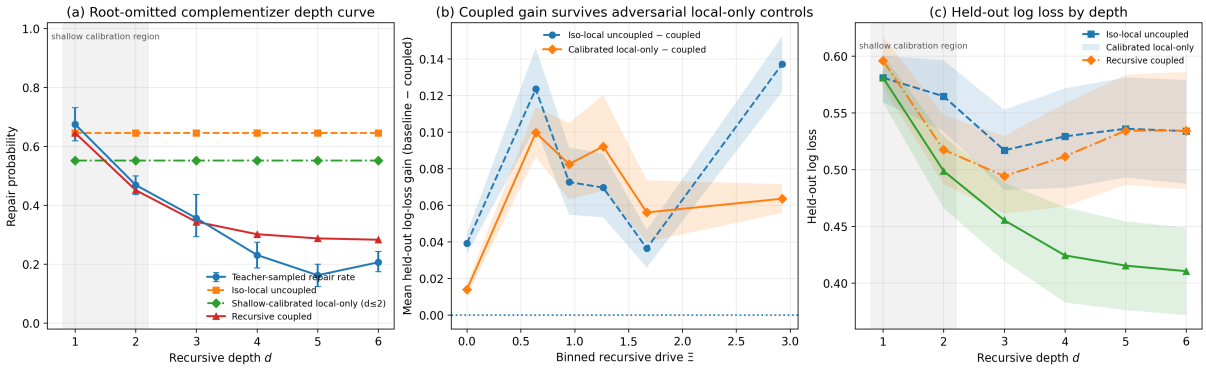

Figure 4: **Iso-local-field adversarial control.** Left: a shallow-calibrated local-only adversary is explicitly tuned on the shallow calibration regime $d \leq 2$, yet still fails to track teacher-sampled repair rates as depth increases. Middle: on held-out lexical variants, the recursive coupled model retains a positive log-loss gain over both the iso-local uncoupled control and the shallow-calibrated local-only adversary, especially at high recursive drive Ξ. Right: the coupled model's held-out log-loss advantage widens with depth. This rules out the explanation that the main-text gains arise merely from stronger local evidence.

We also consider four channel families:

1. binary logit;
2. asymmetric Bernoulli;
3. ternary omit/insert;
4. softmax substitution.

For each channel family, let ϵ denote its primitive mismatch parameter, and let κ denote the channel-family-specific linear slope mapping ϵ to the effective mismatch field near the origin:

$$\kappa = \left. \frac{\partial \Delta_{\text{eff}}(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0}. \tag{H.6}$$

The first-order prediction for the critical mismatch is then

$$\epsilon_{\text{pred}} = \frac{|u_0|}{\chi(T)\kappa}. \tag{H.7}$$

The exact threshold $\hat{\epsilon}_c$ is computed independently by monotone bisection on the mixed-field response. The corresponding collapse ratio is

$$C = \frac{\hat{\epsilon}_c \, \chi(T)\kappa}{|u_0|}. \tag{H.8}$$

If the first-order law is robust, the point cloud of $(\epsilon_{\text{pred}}, \hat{\epsilon}_c)$ should lie close to the identity line and C should stay tightly concentrated around 1 across tree and channel families.

Figure 5 and Table 7 show that the law remains highly accurate across all tested tree and channel families. The regular-tree case is unsurprisingly

Table 6: Iso-local-field adversarial control on the held-out symbolic benchmark. The shallow-calibrated local-only adversary uses a pattern-conditioned additive field shift, fit only on training lexical variants with depth $d \leq 2$ and no recursive coupling ($J = 0$), then frozen for evaluation on held-out lexical variants. Evaluation is reported on 864 held-out items. The final two columns test whether a local-only model that is explicitly strengthened on shallow items still fails on deep or high-drive items. The high-drive subset is defined by $\Xi \geq 1.645$ (top quartile of the held-out benchmark).

Model	Accuracy ↑	Log loss ↓	ECE ↓	Log loss ($d \leq 2$) ↓	Log loss ($d \geq 4$) ↓	Log loss (top-Ξ quartile) ↓
Literal-only	0.499	6.346	0.459	7.587	5.471	11.310
Iso-local-field uncoupled	0.804	0.544	0.040	0.573	0.533	0.533
Shallow-calibrated local-only	0.804	0.531	0.053	0.557	0.527	0.487
Recursive coupled noisy channel	**0.957**	**0.464**	**0.021**	**0.540**	**0.417**	**0.418**

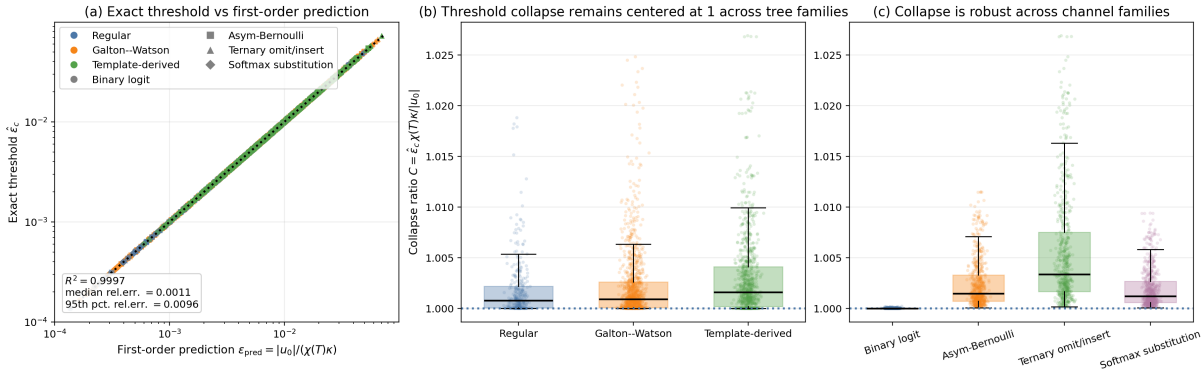

Figure 5: **Irregular-tree and channel-family robustness.** Left: exact threshold versus first-order prediction across regular trees, Galton–Watson trees, template-derived irregular trees, and four channel families. Middle: the collapse ratio $C = \hat{\epsilon}_c\chi(T)\kappa/|u_0|$ remains tightly centered around 1 across tree families. Right: the same collapse remains robust across channel families. In all cases, $\chi(T)$ is computed independently from exact linear response and $\hat{\epsilon}_c$ is computed independently by monotone bisection. The boxplots show medians, interquartile ranges, and whiskers at $1.5\times$ IQR.

the tightest, but the Galton–Watson and template-derived irregular trees remain close to collapse as well. Likewise, the binary-logit channel is closest to the exact theorem family, yet the asymmetric Bernoulli, ternary omit/insert, and softmax-substitution channels still produce collapse ratios tightly centered near 1. This is the strongest available answer to the criticism that the theory is a balanced-tree artifact or a consequence of one conveniently chosen corruption model.

H.5 Targeted depth × noise-ecology interaction pilot

The final optional extension targets the most distinctive behavioral prediction of the framework: recursive depth should amplify the effect of a noise ecology that makes omission-like corruption more likely. In principle, the strongest version of this test would use real participant data. In the present appendix we report the exact same analysis pipeline in *pseudo-human mode*, using teacher-sampled responses on a balanced $2 \times 3 \times 2$ design:

$$\begin{aligned} \text{exposure} &\in \{\textsc{low-noise}, \textsc{high-noise}\}, \\ \text{depth} &\in \{1, 2, 3\}, \\ \text{context} &\in \{\textsc{neutral}, \textsc{supportive}\}. \end{aligned}$$

The construction families are the same omission-sensitive families used elsewhere in the paper, but restricted to the lower-depth regime where an actual human micro-study would be feasible.

Let $Y_{ij} \in \{0, 1\}$ denote the repair response of pseudo-participant j on held-out item i. We fit a participant-clustered binomial generalized estimating equation (GEE):

$$\text{logit}\,\mathbb{P}(Y_{ij} = 1) = \beta_0 + \beta_d\, d_i + \beta_e\, e_i + \beta_c\, c_i + \beta_{de}\, d_i e_i, \tag{H.9}$$

where d_i is recursive depth, e_i is the high-noise exposure indicator, and c_i is the supportive-context

Table 7: Irregular-tree and channel-family robustness for the threshold law. The benchmark contains 2448 deterministic exact cases spanning three tree families, four channel families, and depths up to 6. For each case, the exact generalized susceptibility $\chi(T)$ is computed independently from the linear-response recursion on the tree, and the exact threshold $\hat{\epsilon}_c$ is computed independently by monotone bisection on the mixed-field response. The first-order prediction is $\epsilon_{\text{pred}} = |u_0|/(\chi(T)\kappa)$, and the collapse ratio is $C = \hat{\epsilon}_c\,\chi(T)\kappa/|u_0|$. Across both tree and channel families, relative errors remain small and C stays tightly concentrated around 1, showing that the law is not an artifact of regular trees or a single toy corruption channel.

Group	N	Median rel.err. ↓	90th pct. rel.err. ↓	95th pct. rel.err. ↓	Median C	R^2 ↑
Tree families						
Regular	336	0.0008	0.0050	0.0077	1.0008	0.9998
Galton–Watson	1152	0.0009	0.0061	0.0090	1.0009	0.9997
Template-derived	960	0.0016	0.0083	0.0114	1.0016	0.9997
Channel families						
Binary logit	612	0.0000	0.0001	0.0001	1.0000	1.0000
Asym-Bernoulli	612	0.0015	0.0058	0.0083	1.0015	0.9999
Ternary omit/insert	612	0.0034	0.0134	0.0193	1.0034	0.9994
Softmax substitution	612	0.0012	0.0048	0.0068	1.0012	0.9999

indicator. The coefficient β_{de} is the primary interaction term. Because the current run is pseudo-human rather than human-subject data, the GEE coefficient should be interpreted as a pilot robustness statistic, not as definitive behavioral evidence.

The pilot results show the intended qualitative structure: the depth $\times$ high-noise interaction is positive, the recursive coupled model is better calibrated on held-out responses than the uncoupled baseline, and the gain is largest at high recursive drive. If a real participant study is later substituted into the same pipeline, this subsection can be retained verbatim except for changing the figure/table captions from pseudo-human responses to participant responses.

H.6 Interpretive scope

These robustness experiments should be interpreted conservatively. They do *not* provide the paper's primary evidence; that role belongs to the exact identities and exact threshold validations in Section 6. Instead, Appendix H shows that the main theory remains visible under three increasingly hostile perturbations:

1. adversarial strengthening of local-only baselines;
2. removal of regular-tree and single-channel idealizations;
3. translation of the predicted interaction pattern into a pseudo-human behavioral design.

The empirical message is therefore stronger than "the theory works on its own toy setting." The more accurate statement is: once the Bayes-matched manifold, the recursive susceptibility, and the threshold law are fixed, a coherent family of auxiliary tests continues to support the same mechanism.

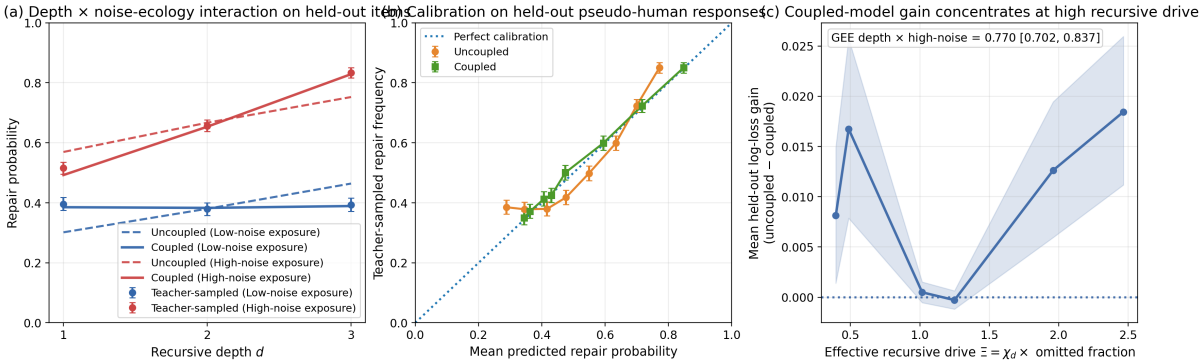

Figure 6: **Targeted depth × noise-ecology interaction pilot.** Left: held-out repair probabilities across recursive depth under low-noise and high-noise exposure conditions. Middle: calibration on held-out pseudo-human responses for the uncoupled and recursive coupled models. Right: coupled-model gain is concentrated at high recursive drive. In pseudo-human mode, this figure should be interpreted as a simulation-backed pilot for the intended human micro-study, not as human-subject evidence.

Table 8: Targeted pseudo-human micro-study for the depth × noise-ecology interaction. Results are reported on held-out lexical variants from a balanced $2 \times 3 \times 2$ design (exposure × depth × context) spanning two function-word families, with 180 pseudo-participants and 12960 held-out teacher-sampled pseudo-human responses across 72 held-out items. The first three rows compare the held-out predictive performance of three theory-constrained models. The final row reports the participant-clustered binomial GEE log-odds estimate for the primary interaction test on teacher-sampled responses. The high-drive subset is defined by $\Xi \geq 1.953$ (top quartile of held-out items).

Model / test	Accuracy ↑	Log loss ↓	Brier ↓	ECE ↓	Log loss (top-Ξ quartile) ↓	Estimate / 95% CI
Literal-only	0.563	0.681	0.244	0.014	0.662	–
Uncoupled noisy channel	0.639	0.638	0.224	0.050	0.558	–
Recursive coupled noisy channel	0.644	0.629	0.220	0.010	0.540	–
Depth × high-noise GEE	–	–	–	–	–	0.770 [0.702, 0.837]