Unit-Abductive Local Bilinear Models: Closed-Form Prediction from Stable Unit Beliefs

Heyang Gong

Preprint with preliminary diagnostics prepared July 22, 2026

Scope and evidence status. This is a predictive-modeling theory paper with an internally frozen, small-sample empirical diagnostic. It contains no tuned benchmark and makes no general predictive-superiority claim. The Unit Selection Variable U is latent, while $Q_\phi(\mathrm{d}u \mid O)$ is a learner-side unit belief formed from evidence; it is not automatically a calibrated Bayesian posterior or a recovered physical identity. The paper does not introduce treatment, intervention, potential-outcome, counterfactual, or causal-identification semantics. "Local" means unit-conditioned affine prediction, not top-k retrieval, neighborhood-weighted fitting, prototype routing, or an inference-time local solve.

Abstract

We introduce a predictive model organized around a latent Unit Selection Variable whose values encode candidate unit identities or states through a response-relevant representation. A sample is an observed event that supplies evidence about that unit; the learner forms a unit belief, which then modulates a linear predictor. For unit candidate u, the slope and intercept are affine, $w(u) = \beta + Bu$ and $b(u) = \alpha_0 + a^\top u$, yielding the bilinear response $Y = \alpha_0 + \beta^\top X + a^\top U + X^\top BU + \sigma E$. We call the model local because fixing a candidate unit gives an affine view of the predictor–response relation; evidence-specific location and scale determine which such views carry predictive mass. A centered Taylor result shows how this form retains the mixed input–unit curvature of a smooth response surface while exposing the omitted pure curvatures and third-order remainder. If the unit-belief coordinates and event noise are independent symmetric stable variables, we derive the marginal predictive law in closed form at the level of its characteristic function and scale. The Cauchy specialization has an elementary likelihood, quantiles, and bounded location and log-scale scores. The matched Gaussian endpoint has variance twice the squared stable scale. We also prove log-score Fisher consistency for the observable response law, characterize an exact varying-coefficient reduction, and exhibit latent-coordinate and unit/event-scale non-identifiability. Closed form refers to marginalization and the endpoint likelihoods; model parameters are still learned by numerical maximum likelihood. The result is a compact prediction-only account of global nonlinearity as a shared family of unit-belief-conditioned linear views. An initial three-dataset screen generated the hypothesis that the Cauchy endpoint is less sensitive than its matched Gaussian endpoint to gross development-label outliers. Before inspecting results on nine additional regression datasets, we internally froze the panel, five seeds, corruption arms, model budgets, and a dataset-level progression rule; this was not an external preregistration. Under 10% balanced $\pm 5 s_y$ contamination, all nine extension datasets met the rule that both median normalized degradation and absolute dirty-RMSE gap favor Cauchy; their cross-dataset medians were -0.160 and -0.218 clean-training target-SD units. Under exact 20% value-changing label shuffle, only four of nine met the same descriptive criterion. This reaches a suite-specific internal progression gate for severity and loss-only ablations, not a general robustness or component-isolation claim.

1 Introduction

A nonlinear predictor can be understood either as one opaque global map or as a structured family of simpler views. Local polynomial methods take the second route by fitting an input-neighborhood approximation [4, 7]. Mixtures of experts route examples among learned predictors [15, 16]; varying-coefficient models let observed variables alter regression parameters [8, 14]. These views are useful, but they do not distinguish evidence used to characterize an observation-specific unit from the predictor at which the response is evaluated.

We study that distinction directly. Let O denote evidence, X a predictor, Y a scalar prediction target, and U a latent Unit Selection Variable. The variable U is ontology-level: its values distinguish candidate units or unit states and carry their response-relevant representation. An observed row is an event or record associated with a unit; its row index is not the unit identity, and the same unit may in principle support multiple observations. The learner maps evidence to a probability kernel $Q_\phi(\,\mathrm{d}u \mid O)$, called a *unit belief*. The response coefficients are affine functions of a candidate u:

$$w(u) = \beta + Bu, \qquad b(u) = \alpha_0 + a^\top u. \tag{1}$$

For fixed u, prediction is affine in x; across candidates, slopes and intercepts vary through one shared low-rank parameterization. Integrating the candidate-conditioned predictor over Q_ϕ turns uncertainty about the selected unit into a predictive distribution.

The word *local* is intentionally narrower than a universal claim that a nonlinear function is exactly linear. A differentiable surface has a local Taylor expansion, and its mixed input–unit Hessian produces precisely a bilinear term. Pure input and pure unit curvature remain approximation error. Operationally, locality in this paper is unit-conditioned: evidence determines a location–scale region in unit space, and every unit candidate in that region indexes an affine $x \mapsto y$ view. There is no reference bank, nearest-neighbor stage, or query-specific optimizer.

Symmetric stable beliefs make the composition analytic. Cauchy beliefs are the main case: their location and scale define a heavy-tailed candidate geometry, and an affine combination remains Cauchy. A Gaussian model is not introduced as an unrelated alternative; it is the $\rho = 2$ endpoint of the same $\mathrm{S}_\rho\mathrm{S}$ characteristic-function convention. This matching matters: the Gaussian standard deviation is $\sqrt{2}$ times its stable scale. For general ρ, propagation has a closed-form characteristic function and scale, although an elementary density is generally unavailable [22, 29].

Contributions.

- We define Unit-Abductive Local Bilinear Models (UALBMs), separating evidence O, predictor X, Unit Selection Variable U, learner-side unit belief Q_ϕ, and event noise E in a factual-prediction model.
- We give a centered local-bilinear Taylor theorem with an explicit error decomposition: retained mixed curvature, omitted pure curvatures, and a controlled third-order remainder.
- We prove exact marginal propagation for independent symmetric stable unit-belief coordinates and event noise, then derive elementary Cauchy and matched Gaussian likelihoods and quantiles.
- We identify the observable varying-coefficient reduction, prove response-law Fisher consistency of the log score, and state constructive latent-coordinate and unit/event-scale non-identifiability results.

- We report an audited preliminary screen against Direct MLP and a fixed-default gradient-boosting reference. It isolates a narrow Cauchy signal under gross additive label outliers, mixed behavior under label shuffling, and a concrete bilinear extrapolation failure.

The mathematical ingredients—Taylor expansion, bilinear regression, stable closure, and proper log scoring—are individually classical. The claim is the specific predictive composition and its clarified boundaries, not a new universal representation theorem. If the composition offers no distinction beyond standard varying-coefficient or distributional regression in a target application, it should be treated as a technical parameterization rather than a new predictive paradigm.

2 Predictive setup and unit belief

2.1 Variables and observation regime

We observe

$$\mathcal{D}_n = \{(o_i, x_i, y_i)\}_{i=1}^n, \qquad o_i \in \mathcal{O},\ x_i \in \mathbb{R}^p,\ y_i \in \mathbb{R}. \tag{2}$$

The roles of O and X are distinct: O supplies evidence about the selected unit, whereas X is the argument of the predictor. An application may encode the same observed vector into both roles, but equality of values does not collapse the two interfaces.

Definition 1 (Unit Selection Variable and unit belief)**.** The Unit Selection Variable is an ontology-level latent model variable $U \in \mathbb{R}^d$. Its values distinguish candidate units or unit states through a response-relevant representation. A sample row is an observed event or record associated with a unit; it carries evidence about U but is not itself the unit. A unit belief is a learned Markov kernel

$$Q_\phi : \mathcal{O} \times \mathcal{B}(\mathbb{R}^d) \to [0,1], \qquad Q_\phi(A \mid o) = \Pr_\phi(\widetilde{U} \in A \mid O = o), \tag{3}$$

where $\widetilde{U}$ denotes a learner-side computational candidate. The kernel is the learner's belief about candidate values of U after seeing evidence. It need not arise from a population prior or evidence likelihood, and is not called a calibrated posterior without an additional calibration argument.

The tilde in Eq. (3) prevents a semantic error: drawing a computational candidate from a belief is not a claim that physical identity is redrawn. For prediction, only the composed response law is required.

Unit selection is not sample selection. The term *selection* in U first names an ontological and inferential question: which candidate value of the Unit Selection Variable characterizes the unit associated with an observed event, and what response-relevant representation does that value carry? After seeing evidence $O = o$, the learner's uncertain answer is $Q_\phi(\,\mathrm{d}u \mid o)$. Only after a candidate u passes through $u \mapsto \{w(u), b(u)\}$ does it induce a local coefficient view; unit selection is therefore not defined as routing among pre-existing local views. Marginalization over Q_ϕ is the computational consequence inside that row's predictive law. A sample-inclusion indicator $R \in \{0,1\}$, if introduced, would instead describe whether a row enters the source sample, or how source rows should be weighted to represent a target population. Thus a unit belief is neither an inclusion propensity nor an importance weight. The baseline model in this paper uses the ordinary unweighted supervised sample average; it does not fit a sample-selection propensity or claim target-population transport.

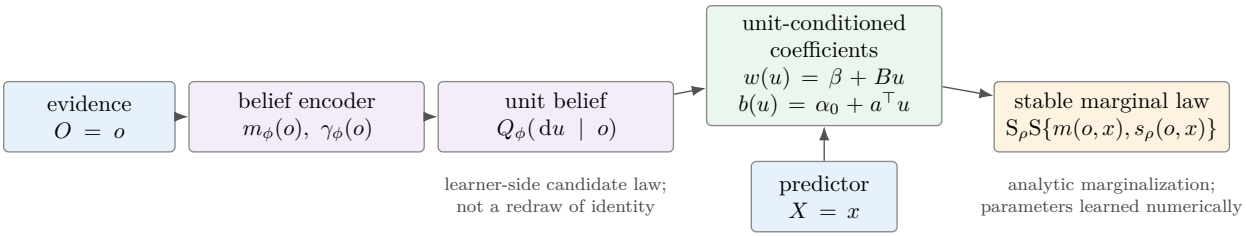

Figure 1: UALBM prediction pipeline. Evidence forms a unit belief; candidates index a shared affine coefficient family; the predictor is evaluated under those coefficients; and stable closure gives the marginal response law. No retrieval, prototype assignment, or local optimization occurs at inference.

2.2 Stable unit beliefs

For $0 < \rho \le 2$, write $Z \sim \mathrm{S}_\rho\mathrm{S}(m, \gamma)$ when

$$\mathbb{E}[e^{itZ}] = \exp\{imt - |\gamma t|^\rho\}, \qquad m \in \mathbb{R}, \quad \gamma > 0. \tag{4}$$

Thus $\mathrm{S}_1\mathrm{S}(m, \gamma) = \mathrm{Cauchy}(m, \gamma)$ and $\mathrm{S}_2\mathrm{S}(m, \gamma) = \mathcal{N}(m, 2\gamma^2)$. We use ρ for the stability index and α_0 for the scalar intercept.

The encoder emits locations $m_\phi(o) \in \mathbb{R}^d$ and positive scales $\gamma_\phi(o) \in \mathbb{R}^d_{>0}$, defining the factorized belief

$$Q_\phi(\mathrm{d}u \mid o) = \bigotimes_{j=1}^{d} \mathrm{S}_\rho\mathrm{S}\{\mathrm{d}u_j; m_{\phi,j}(o), \gamma_{\phi,j}(o)\}. \tag{5}$$

The Cauchy location $m_{\phi,j}$ is not a mean and $\gamma_{\phi,j}$ is not a standard deviation; neither Cauchy mean nor variance exists. Positivity can be enforced by $\gamma_\phi(o) = \gamma_{\min} + \mathrm{softplus}\{h_{\phi,\mathrm{scale}}(o)\}$.

2.3 Unit-abductive bilinear predictor

Let $B \in \mathbb{R}^{p\times d}$, $\beta \in \mathbb{R}^p$, $a \in \mathbb{R}^d$, and $\alpha_0 \in \mathbb{R}$. The unit-conditioned slope and intercept are Eq. (1), and the shared candidate response model is

$$Y = b(\widetilde{U}) + w(\widetilde{U})^\top X + \sigma E \tag{6}$$

$$= \alpha_0 + \beta^\top X + a^\top \widetilde{U} + X^\top B\widetilde{U} + \sigma E, \qquad \sigma > 0, \tag{7}$$

where $E \sim \mathrm{S}_\rho\mathrm{S}(0, 1)$ is independent of the computational candidate. Define

$$g(x) = a + B^\top x \in \mathbb{R}^d. \tag{8}$$

For fixed x, the random component is $g(x)^\top \widetilde{U} + \sigma E$; for fixed u, $x \mapsto b(u) + w(u)^\top x$ is affine.

3 Unit-conditioned local bilinearity

3.1 Operational meaning of local

Definition 2 (Unit-conditioned local view)**.** For evidence o, a unit-conditioned local view is the family

$$\mathcal{L}(o) = \{x \mapsto b(u) + w(u)^\top x : u \text{ lies in a declared central region of } Q_\phi(\cdot \mid o)\}. \tag{9}$$

Every member is affine in x. The location and scale of the unit belief determine the center and breadth of the view; they do not define a neighborhood of training rows.

For independent Cauchy coordinates, a concrete central level-r region is

$$\mathcal{C}_r(o) = \prod_{j=1}^{d} \left[m_{\phi,j}(o) - \gamma_{\phi,j}(o) \tan\frac{\pi r}{2}, m_{\phi,j}(o) + \gamma_{\phi,j}(o) \tan\frac{\pi r}{2} \right], \qquad 0 < r < 1. \tag{10}$$

Each coordinate interval has Cauchy probability r, so independence gives $Q_\phi\{\mathcal{C}_r(o) \mid o\} = r^d$. This definition uses quantiles, not moments.

3.2 Centered Taylor motivation

The following result states exactly what a local bilinear patch keeps and what it discards. It is a motivation for the inductive bias, not a claim that an arbitrary global surface equals one bilinear polynomial.

Theorem 3 (Centered local-bilinear approximation)**.** *Let $f : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ be three times continuously differentiable on a convex neighborhood containing the segment between (x_0, u_0) and (x, u). Set $h = x - x_0$, $k = u - u_0$, and let $H_{xu} = \nabla^2_{xu} f(x_0, u_0)$. Define*

$$T_{\mathrm{BL}}(x, u) = f_0 + \nabla_x f_0^\top h + \nabla_u f_0^\top k + h^\top H_{xu} k. \tag{11}$$

Then

$$f(x, u) - T_{\mathrm{BL}}(x, u) = \frac{1}{2} h^\top H_{xx} h + \frac{1}{2} k^\top H_{uu} k + R_3, \tag{12}$$

$$|R_3| \le \frac{M_3}{6} \|(h, k)\|_2^3, \tag{13}$$

where all displayed Hessians are evaluated at (x_0, u_0) and M_3 bounds the operator norm of the third derivative along the segment. Consequently,

$$|f - T_{\mathrm{BL}}| \le \frac{1}{2} \|H_{xx}\|_{\mathrm{op}} \|h\|_2^2 + \frac{1}{2} \|H_{uu}\|_{\mathrm{op}} \|k\|_2^2 + \frac{M_3}{6} \|(h, k)\|_2^3. \tag{14}$$

Moreover, T_{BL} can be written uniquely in the uncentered form $\alpha_0 + \beta^\top x + a^\top u + x^\top B u$ after its coefficients are fixed at (x_0, u_0).

The proof and coefficient conversion are in Appendix B. The mixed term $h^\top H_{xu} k$ is retained; pure x and pure u curvature are explicit error. A richer model could use multiple patches, but top-k fitting, mixture routing, and prototype selection are outside the base UALBM model.

4 Stable marginal prediction

Theorem 4 (Symmetric stable unit-belief propagation)**.** *Suppose Eqs. (5) and (7) hold for one $0 < \rho \le 2$, the candidate coordinates are conditionally independent given $O = o$, and $E \sim \mathrm{S}_\rho\mathrm{S}(0, 1)$ is independent of them. Then*

$$Y \mid O = o, X = x \sim \mathrm{S}_\rho\mathrm{S}\{m(o, x), s_\rho(o, x)\}, \tag{15}$$

with

$$m(o, x) = \alpha_0 + \beta^\top x + g(x)^\top m_\phi(o), \tag{16}$$

$$s_\rho(o, x) = \left[\sigma^\rho + \sum_{j=1}^{d} |g_j(x)|^\rho \gamma_{\phi,j}(o)^\rho \right]^{1/\rho}. \tag{17}$$

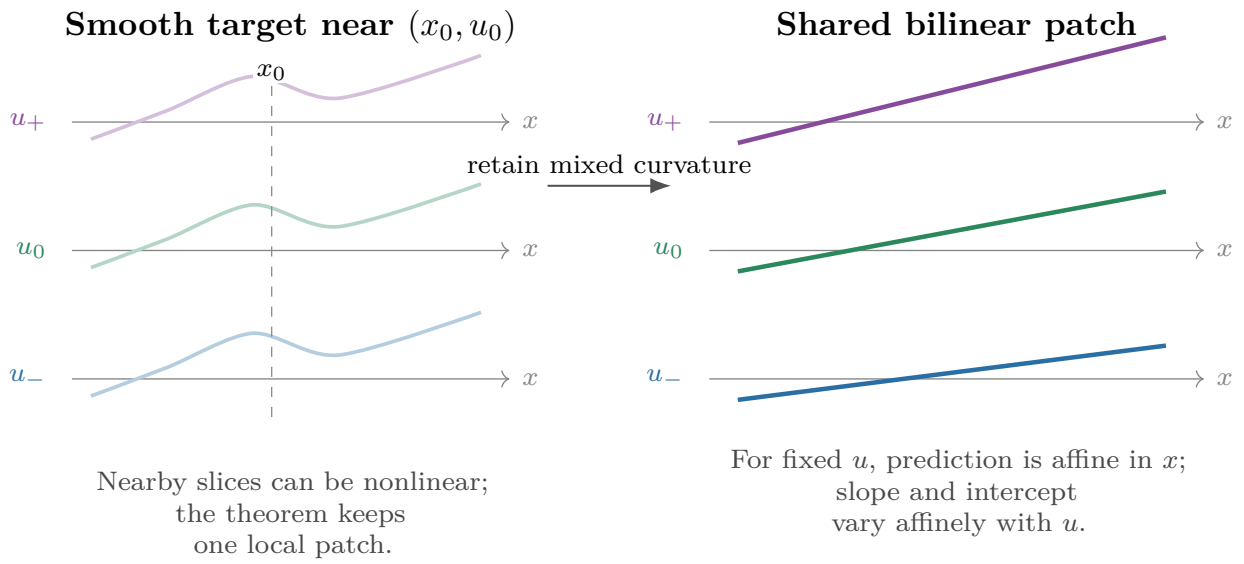

Figure 2: Taylor motivation versus model definition. The left panel represents nearby slices of an unknown smooth surface. The right panel is the centered bilinear patch: each fixed-unit slice is affine, and the mixed Hessian controls how its slope changes with unit state. Rows are candidate unit states, not neighboring observations; no input-neighborhood fitting is implied.

Equivalently, the conditional characteristic function is

$$\varphi_{Y|o,x}(t) = \exp\{itm(o,x) - |s_\rho(o,x)t|^\rho\}. \tag{18}$$

The characteristic-function proof is in Appendix C. It is exact; no Monte Carlo candidates are needed. For generic ρ, Eq. (18) is a closed-form law specification, but the density generally requires numerical evaluation.

4.1 Cauchy main model

Corollary 5 (Cauchy response law, likelihood, and quantiles)**.** *At $\rho = 1$,*

$$Y \mid o, x \sim \text{Cauchy}\{m(o,x), s_1(o,x)\}, \qquad s_1(o,x) = \sigma + \sum_{j=1}^{d} |g_j(x)|\gamma_{\phi,j}(o). \tag{19}$$

For residual $r = y - m$ and $s = s_1(o,x)$, the negative log likelihood is

$$\ell_{\text{C}}(y; m, s) = \log(\pi s) + \log\left\{1 + \left(\frac{r}{s}\right)^2\right\}. \tag{20}$$

Its qth predictive quantile is

$$F^{-1}(q) = m + s\tan\{\pi(q - 1/2)\}, \qquad 0 < q < 1. \tag{21}$$

The location and log-scale derivatives satisfy

$$\left|\frac{\partial \ell_{\text{C}}}{\partial m}\right| \le \frac{1}{s}, \qquad \left|\frac{\partial \ell_{\text{C}}}{\partial \log s}\right| \le 1. \tag{22}$$

The bounded scores limit the influence of arbitrarily large residuals on the location and log-scale outputs for fixed positive s. This analytic property does not by itself imply empirical robustness; Section 6 reports only a narrow exploratory contamination signal. Cauchy prediction is summarized by location, scale, CDF, and quantiles—never by a nonexistent conditional mean or variance.

4.2 Matched Gaussian endpoint

Corollary 6 (Gaussian endpoint under the stable convention)**.** *At $\rho = 2$,*

$$s_2(o, x) = \left[\sigma^2 + \sum_{j=1}^{d} g_j(x)^2 \gamma_{\phi,j}(o)^2\right]^{1/2}, \tag{23}$$

and

$$Y \mid o, x \sim \mathcal{N}\{m(o, x), 2s_2(o, x)^2\}. \tag{24}$$

The matched Gaussian NLL is

$$\ell_{\mathrm{G}}(y; m, s_2) = \log(2\sqrt{\pi}\, s_2) + \frac{(y - m)^2}{4s_2^2}. \tag{25}$$

If an implementation emits conventional standard deviations $\tau_j(o)$ and σ_{N}, the exact conversion is $\gamma_j = \tau_j/\sqrt{2}$ and $\sigma = \sigma_{\mathrm{N}}/\sqrt{2}$.

Thus the Cauchy and Gaussian models share the encoder, coefficient family, stable scale convention, and marginalization argument. Replacing s_2 by a Gaussian standard deviation without the $\sqrt{2}$ conversion would not be a matched baseline.

4.3 Three neural realizations

For the controlled diagnostic in Section 6, we distinguish the three architectures in Fig. 3. All three use the same evidence-to-representation architectural template—a perception MLP followed by a linear hidden head—but they own independent parameters. The Direct MLP combines its deterministic representation $r_D(o)$ with the predictor x in a scalar output head and is trained by point-prediction loss. It defines no unit belief, stable scale, or analytic marginal response law.

The Gaussian and Cauchy UALBM lanes are endpoint-matched. Each adds the same positive stable-scale head, factorized unit-belief layer, and globally learned bilinear marginalizer $(\alpha_0, \beta, a, B, \sigma)$. Their topology and latent dimension can be identical; they differ in the common stability index assigned to both U and E, the induced scale aggregation, and the endpoint negative log likelihood. Thus Gaussian versus Cauchy is a matched stable-endpoint comparison, whereas Direct MLP versus either distributional model is an algorithm-level baseline rather than a single-factor ablation.

5 Learning and observable content

5.1 Numerical maximum likelihood

For $\theta = (\phi, \alpha_0, \beta, a, B, \sigma)$, UALBM minimizes

$$\widehat{\mathcal{Q}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\{y_i; m_\theta(o_i, x_i), s_{\rho,\theta}(o_i, x_i)\}, \tag{26}$$

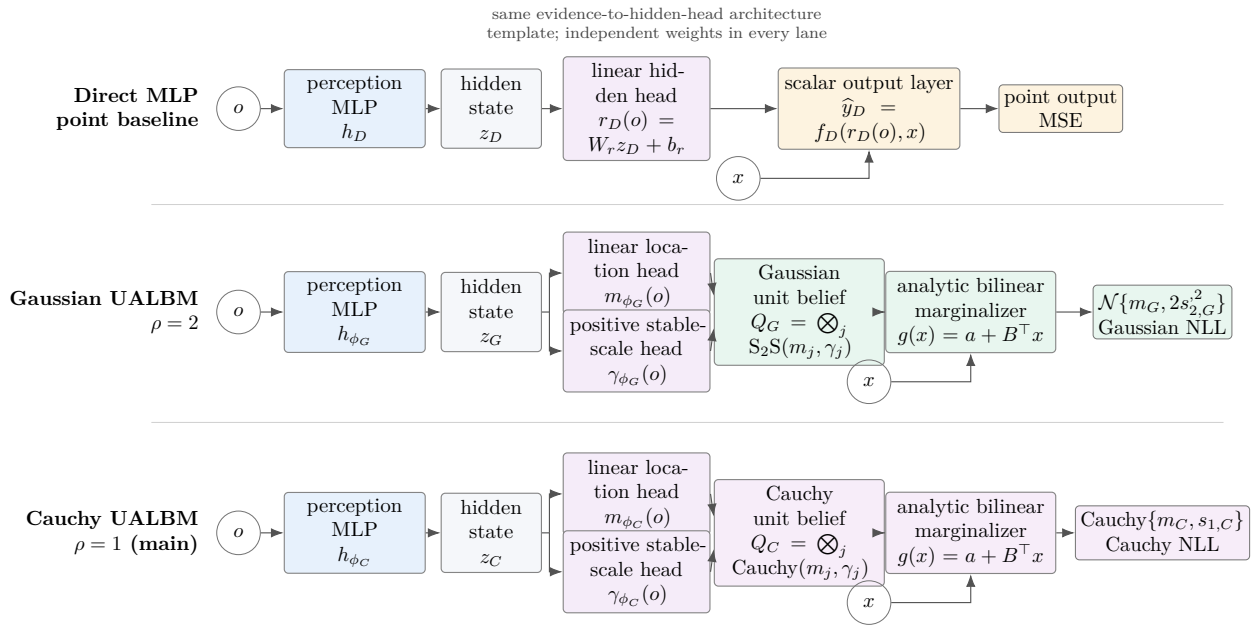

Figure 3: Three neural realizations used in the preliminary screen. The lanes reuse the same evidence-encoder and linear-hidden-head architecture template but train independent parameters. Direct MLP combines a deterministic representation with x in a scalar point-prediction head and defines no unit belief. The two UALBM lanes add the same positive stable-scale head and analytic bilinear marginalizer, differing in ρ, scale aggregation, and endpoint NLL. Under our convention the Gaussian response is $\mathcal{N}\{m_G, 2s^2_{2,G}\}$, so its standard deviation is $\sqrt{2}\, s_{2,G}$; Cauchy quantities are location and scale, never mean or standard deviation. Gaussian versus Cauchy is the endpoint-matched comparison; Direct MLP remains an algorithm-level point baseline.

using the elementary Cauchy or Gaussian NLL. "Closed-form prediction" means that the candidate integral and endpoint likelihood are analytic. It does not mean that the neural encoder or bilinear parameters have a closed-form maximum likelihood estimator. They are optimized numerically. The analytic forward pass costs $O(pd + d)$ per row after encoding and requires no latent samples; details and gradients are in Appendix E.

Theorem 7 (Log-score Fisher consistency for the response law)**.** *Let $r_0(y \mid o, x)$ be the true conditional density and let a dominated candidate class contain r_0. If the relevant log risks are finite, every population minimizer of*

$$\mathcal{Q}(r) = \mathbb{E}[-\log r(Y \mid O, X)] \tag{27}$$

agrees with $r_0(\cdot \mid O, X)$ almost surely. This identifies the observable response law under correct specification, not the latent unit coordinates or their factorization.

5.2 Varying-coefficient reduction

Proposition 8 (Observable varying-coefficient form)**.** *Define*

$$c(o) = \alpha_0 + a^\top m_\phi(o), \qquad v(o) = \beta + B m_\phi(o). \tag{28}$$

Then the predictive location is exactly

$$m(o, x) = c(o) + v(o)^\top x, \qquad v(o) \in \beta + \mathrm{col}(B). \tag{29}$$

Consequently, location data can at most identify the evidence-varying intercept and slope and, under sufficient support, their affine slope subspace. They do not select a unique m_ϕ, a, and B factorization.

In the Cauchy model, $m(o, x)$ is a response *location*, not a conditional mean. The reduction exposes UALBM's closest classical neighbor and its main observable structural restriction: evidence-indexed slopes lie in a shared affine subspace of dimension at most $\operatorname{rank}(B)$.

5.3 Non-identifiability boundaries

Proposition 9 (Latent affine reparameterization)**.** *Let D be an invertible monomial matrix (a permutation times a nonsingular diagonal matrix) and $c_U \in \mathbb{R}^d$. Replacing the candidate coordinate by $V = D\widetilde{U} + c_U$, pushing the unit belief forward, and transforming (α_0, β, a, B) as specified in Appendix D.3 leaves the complete predictive law unchanged for every (o, x). Permutations and sign changes are included. Therefore numerical unit coordinates are not identified by the response law.*

Proposition 10 (Unit/event Cauchy scale split is not identified)**.** *The scalar submodel*

$$\widetilde{U} \mid o \sim \text{Cauchy}\{m_\phi(o), \gamma_\phi(o)\}, \qquad E \sim \text{Cauchy}(0, 1), \qquad Y = \widetilde{U} + \sigma E \tag{30}$$

has response scale $\gamma_\phi(o) + \sigma$. For every constant $0 < \delta < \sigma$, replacing $\gamma_\phi(o)$ by $\gamma_\phi(o) + \delta$ and σ by $\sigma - \delta$ leaves all conditional response laws unchanged. Hence labels such as "unit-belief scale" and "event scale" are modeling assignments unless additional observations or restrictions separate them.

These results do not make the predictive distribution vacuous: the marginal location, scale, quantiles, and log score remain observable targets. They do rule out recovery claims based only on a fitted latent factorization.

6 Empirical diagnostics: discovery and prospective extension

Question and study chronology. We ask whether the three realizations in Fig. 3 function as ordinary supervised predictors and whether the matched Cauchy endpoint is less sensitive than the Gaussian endpoint to corrupted development labels. A first screen on California Housing, Wine Quality, and Abalone forms the prior-discovery evidence and generated a gross-outlier hypothesis. Before running any model on nine additional regression datasets, we internally froze the extension panel, five seeds, corruptions, model budgets, and dataset-level progression rule. This temporal partition is not an external preregistration, and the study remains a small-sample diagnostic rather than a tuned benchmark.

Protocol. The discovery datasets are California Housing [23], Wine Quality [5], and Abalone [21]. The frozen extension comprises Diabetes [6], Concrete Compressive Strength [31], Energy Efficiency (heating load) [30], Airfoil Self-Noise [3], Yacht Hydrodynamics [10], Auto MPG [26], QSAR Fish Toxicity [1], Protein Tertiary Structure [27], and Superconductivity [13]; provenance and preprocessing are recorded in Appendix F. For each dataset and seed in $\{42, 43, 44, 45, 46\}$, we use at most 500 rows, split them 60/20/20 into train/validation/test partitions, and fit feature standardization on training rows only. The observed interface is collapsed to $O = X$ numerically while the two model arguments remain separate. Each neural method uses a two-layer SiLU encoder of width 64, an eight-dimensional hidden or unit representation, Adam, at most 80 epochs, and dirty-validation early stopping. Direct MLP minimizes MSE. Gaussian and Cauchy UALBM minimize their matched endpoint NLLs. The contextual tree reference is fixed-default

`GradientBoostingRegressor` from scikit-learn [25]; it is untuned and not topology matched. The complete matrix has $12 \times 5 \times 3 \times 4 = 720$ separately trained result rows. Full hyperparameters and audit rules are in Appendix F.

The contamination screen independently corrupts train and validation labels while leaving test inputs and labels untouched. The shuffle arm changes the numeric value of exactly 20% of labels, even when target values repeat. The gross-outlier arm selects exactly 10% and adds count-balanced perturbations $\pm 5 s_y$, where s_y is the clean training-label population standard deviation. All methods share one frozen corruption plan for each dataset–seed–arm cell. Neural target normalization is refitted on the current arm's training labels, so the screen evaluates the complete dirty-development pipeline rather than a loss-only ablation.

Implementation oracle and integrity checks. Before reading predictive results, a float64 oracle checked a hand-computed bilinear case, analytic stable propagation, automatic gradients, singleton/batch equivalence, and independent latent composition. The maximum gradient discrepancy was 1.22×10^{-15}; with 50,000 latent draws, the Gaussian and Cauchy probability-integral-transform KS distances were 0.00451 and 0.00388. All 720 outputs were finite. Hash-based checks verified disjoint row identities within each split, clean test labels, exact corruption plans, and identical corrupted labels across methods in every matched cell.

Prior-discovery clean prediction. Table 1 reports the median clean-test RMSE across the five paired seeds for the three original datasets. No method dominates clean prediction. Both UALBM endpoints function as ordinary predictors, but this screen provides no evidence of a uniform accuracy advantage over Direct MLP or gradient boosting.

Table 1: Prior-discovery median clean-test RMSE across five paired seeds. Each screen uses 500 rows and a 60/20/20 split. Gradient boosting is an untuned fixed-default reference, not a hyperparameter-matched competitor.

Dataset	Direct MLP	Gaussian	Cauchy	Gradient boost
California	0.6777	0.7528	0.6976	0.6239
Wine	0.7818	0.7975	0.8053	0.8033
Abalone	2.1743	2.3505	2.3479	2.3070

Prior-discovery contamination screen. For method m and a dirty arm c, define the paired clean-test degradation

$$\Delta_m^{(c)} = \mathrm{RMSE}_{m,c} - \mathrm{RMSE}_{m,\mathrm{clean}} . \tag{31}$$

The matched contrast is $D^{(c)} = \Delta_{\mathrm{C}}^{(c)} - \Delta_{\mathrm{G}}^{(c)}$; $D^{(c)} < 0$ means Cauchy degrades less for that paired seed. We always report degradation together with absolute dirty-arm RMSE so that a weak clean anchor cannot manufacture a robustness conclusion.

Under 10% additive $\pm 5 s_y$ contamination, Cauchy has smaller paired degradation than Gaussian in 14 of 15 runs and lower absolute dirty-arm RMSE in 13 of 15. Under 20% value-changing shuffling, the corresponding counts are 11 of 15 and 8 of 15; on Wine, Cauchy has lower dirty-arm RMSE in only one of five runs. We therefore interpret Table 2 as a narrow gross-outlier signal, not evidence of general label-noise robustness. The endpoint pipelines also differ in stable-scale aggregation, fitted target normalization, and validation selection, so the screen does not isolate the bounded Cauchy score as the sole cause.

Table 2: Prior-discovery matched endpoint contamination screen. Dirty G/C are median clean-test RMSEs in the corrupted development arm; Δ columns are median paired changes from the corresponding clean arm. Win counts are out of five seeds.

Dataset	Corruption	Dirty G	Dirty C	Δ_G	Δ_C	C Δ wins	C dirty wins
California	10% $\pm 5s_y$	1.1044	0.6884	0.1479	0.0004	4/5	5/5
Wine	10% $\pm 5s_y$	0.8920	0.8017	0.0963	-0.0122	5/5	5/5
Abalone	10% $\pm 5s_y$	2.5536	2.3158	0.1945	-0.0322	5/5	3/5
California	20% shuffle	0.8145	0.6515	0.0018	-0.0063	4/5	5/5
Wine	20% shuffle	0.8053	0.8803	0.0298	0.0452	3/5	1/5
Abalone	20% shuffle	2.4168	2.3499	0.0753	0.0109	4/5	2/5

Table 3: Dataset-level extension results. Entries are within-dataset medians over five paired seeds in clean-training target-SD units. D/s_y is the Cauchy-minus-Gaussian degradation contrast; gap/s_y is dirty Cauchy RMSE minus dirty Gaussian RMSE. A dataset supports the direction only when both entries for that arm are negative.

	10% $\pm 5s_y$ outliers		20% shuffle	
Dataset	D/s_y	gap/s_y	D/s_y	gap/s_y
Diabetes	-0.1579	-0.1759	-0.0040	0.0065
Concrete	-0.1601	-0.2477	-0.0042	-0.0810
Energy (heating)	-0.0968	-0.0902	-0.0062	-0.0109
Airfoil	-0.1707	-0.2240	0.0035	-0.0493
Yacht	-0.3193	-0.4634	0.0159	-0.1808
Auto MPG	-0.1258	-0.1143	0.0017	-0.0193
QSAR Fish	-0.2306	-0.2183	-0.0110	-0.0008
Protein	-0.0721	-0.0495	-0.0022	0.0551
Superconductivity	-0.2670	-0.2965	-0.0013	-0.0027
Supportive datasets	9/9		4/9	
Macro dataset median	-0.1601	-0.2183	-0.0022	-0.0109

Internally frozen prospective extension. For each extension dataset and dirty arm, we divide both $D^{(c)}$ and the dirty Cauchy-minus-Gaussian RMSE gap by that seed's clean-training target standard deviation s_y, then take the median over five seeds. A dataset supports the frozen direction only when both medians are negative. The dataset, rather than any one seed run, is the progression unit.

All nine extension datasets meet both conditions under the one frozen gross-outlier regime; 41 of 45 paired dataset–seed cells agree on both directions, while the corresponding MAE medians agree on all nine datasets. The macro values -0.1601 and -0.2183 are target-SD-normalized differences, not percentage improvements. The shuffle control is heterogeneous: only four of nine datasets meet both conditions, and its macro contrasts lie much closer to zero. On clean extension data, Direct MLP has the lowest median RMSE on two datasets and fixed-default gradient boosting on seven; neither UALBM endpoint is a clean winner. Thus the extension broadens a pattern specific to this synthetic gross-outlier setting, not a clean-accuracy or generic label-noise advantage.

Observed extrapolation failure and limits. One California test observation had standardized average occupancy 113.407, whereas the largest absolute training value of that feature was 4.457. In the clean seed-45 run, Gaussian and Cauchy UALBM predicted -201.509 and -17.881 for target 1.546. All values were finite: this is a bilinear extrapolation failure under the collapsed $O = X$ interface, not numerical overflow. It motivates median-first reporting, explicit tail diagnostics, and future controls on out-of-range predictors. The extension uses random row splits; Concrete, QSAR

Fish, Protein, and Superconductivity contain exact duplicate feature vectors across train and test in at least one seed, so this is not a strict unseen-unit generalization test. Several Gaussian runs also reached the 80-epoch budget. The fixed-budget results therefore establish neither optimizer convergence nor a general claim that Cauchy learning is smoother. With only five seeds, an untuned tree, and no $O \neq X$ application, the present section supports neither universal predictive superiority nor a population-level robustness inference.

7 Relationship to existing work

Varying coefficients and random coefficients. Varying-coefficient models allow slopes to change with an observed index [8, 14]; mixed and random-coefficient models represent subject- or group-level deviations from population effects [20]. UALBM reduces exactly to an evidence-varying affine location model in Proposition 8. Its narrower structural choice is to represent that variation by a learned probability kernel over a Unit Selection Variable and to propagate both its location and scale into an analytic response distribution. This does not make its latent coordinates more identifiable than classical random effects.

Mixtures of experts and hypernetworks. Mixtures of local experts learn a gate over several predictors [15, 16]. Hypernetworks generate the parameters of one network from another input [12]. UALBM is related because evidence modulates coefficients, but the base model has neither a finite expert assignment nor a prototype router: it integrates a continuous unit belief through one affine coefficient map. A deterministic point belief would reduce to a small coefficient-generating network; retaining scale in the response law is what distinguishes the distributional version from an ordinary embedding-conditioned regressor.

Neural Processes and amortized latent models. Conditional and Attentive Neural Processes encode a context set of input–output pairs to predict targets [9, 17]. Variational autoencoders specify an explicit generative model, typically $p(z)p_\theta(O \mid z)$, and train an amortized approximate posterior with an evidence lower bound containing reconstruction and prior-KL terms [18]. UALBM instead targets the supervised conditional law $p(Y \mid X, O)$: $Q_\phi(U \mid O)$ modulates the slope and intercept of an affine-in-X response, and the stable candidate is marginalized analytically under predictive NLL. The current model has no likelihood for O, no reconstruction objective, no ELBO, and no per-observation KL to a latent prior. The two families can both use amortized distribution-valued encoders; Gaussian versus Cauchy is therefore not the fundamental distinction. Without a prior, likelihood for O, and calibration theorem, Q_ϕ is not promoted to a Bayesian posterior or function posterior.

Local linear modeling. Local polynomial regression and LOESS fit around input-space neighborhoods [4, 7]. Bottou and Vapnik [2] analyze learning algorithms that defer computation to a query neighborhood. UALBM uses "local" differently: conditioning on a unit candidate yields an affine view, and the belief identifies a region in unit space. There is no top-k operation, distance-weighted sample fit, or inference-time solve. The centered Taylor theorem supplies approximation motivation but does not silently introduce any of those algorithms.

Stable and distributional regression. Stable-law closure and parameterization are classical [22, 29]. Generalized additive models for location, scale, and shape make multiple distribution parameters functions of covariates [28]; structured distributional regression does the same with

additive predictors [19]. UALBM's analytic Cauchy and Gaussian response laws belong near this literature. The specific restriction is that evidence-dependent response parameters arise by marginalizing an affine stable unit belief through a bilinear coefficient family. Stable closure itself is not claimed as new.

Proper scoring. The logarithmic score is strictly proper under standard integrability conditions [11]. Our Fisher-consistency theorem is its conditional-density specialization. Bounded Cauchy location and log-scale scores explain a robustness property of gradients. The complete-pipeline diagnostic in Section 6 does not isolate those scores from stable-scale aggregation, target normalization, or validation selection, so its outlier pattern cannot be attributed to the score alone.

8 Discussion and limitations

Approximation, not identity. The local Taylor result does not say that every nonlinear relationship is globally or exactly bilinear. It retains the mixed input–unit second derivative and bounds omitted pure curvature near a declared center. A single shared bilinear patch may be misspecified when pure curvature is large or when slopes do not vary in a low-dimensional affine subspace.

Unit belief, not recovered identity. $Q_\phi(\mathrm{d}u \mid o)$ is a learner-side predictive object. The same marginal law admits latent coordinate transformations, and unit versus event scale can be observationally inseparable. A fitted location must not be described as the true unit, and a fitted scale must not be described as calibrated epistemic uncertainty without external validation. This non-identification boundary does not erase the ontology: U still specifies which latent unit object the belief is about, while the data determine how strongly its coordinates can be learned.

Stable-family restrictions. Exact propagation assumes conditionally independent stable coordinates with a common stability index and independent stable event noise. Correlated Gaussian beliefs are tractable with a full covariance, but arbitrary dependent Cauchy beliefs require a spectral-measure specification and no longer reduce to the coordinatewise scale sum in Eq. (17). Generic stable densities also lack elementary likelihoods.

Cauchy semantics. The Cauchy has a location and scale, CDF, and quantiles, but no finite mean or variance. Its bounded score is not an automatic robustness theorem. The Gaussian endpoint is matched only after respecting the stable scale convention $\mathrm{sd} = \sqrt{2}\, s_2$.

Prediction-only scope. The model learns $p(y \mid o, x)$ from factual prediction records. It attaches no treatment, intervention, alternative-world, or causal meaning to X and makes no mechanism-recovery claim. Any such interpretation would be a different paper requiring additional data, assumptions, and identification arguments.

Empirical evidence boundary. The nine-dataset extension was internally frozen after the three-dataset discovery screen, but it is not an external preregistration and the convenience suite is not a random sample of regression tasks. Each dataset contributes at most 500 rows and only five seeds. We test one symmetric gross-outlier fraction and magnitude; we do not test natural, asymmetric, heteroscedastic, or adversarial noise. Both training and validation labels are corrupted, target normalization is refitted, and dirty-validation checkpoint selection remains inside the evaluated

pipeline. The result is therefore not a loss-only ablation and cannot attribute the pattern to the bounded Cauchy score.

The collapsed $O = X$ interface, fixed neural budgets, exact-feature duplicates under some random splits, and untuned tree reference further limit the study. The evidence exposes heterogeneous clean accuracy, a fixed-suite gross-outlier pattern, and an out-of-range bilinear failure, but it does not establish general predictive superiority, calibrated uncertainty, or general label-noise robustness. Severity curves, clean-anchored loss-only ablations, larger samples, and grouped or deduplicated splits are the next empirical gates. Analytic tractability and Fisher consistency under correct specification remain distinct from finite-sample performance.

9 Conclusion

UALBM turns evidence into a unit belief, uses candidate units to index a shared affine coefficient family, and marginalizes that family into a predictive response law. The resulting bilinear model preserves a precise mixed-curvature term from a local Taylor view. Symmetric stable beliefs make propagation exact; the Cauchy and Gaussian endpoints give elementary matched likelihoods. The model's observable contribution is a distributional varying-coefficient predictor with an affine slope-subspace restriction. Its latent coordinates and uncertainty decomposition are not generally identified. These positive and negative results together define a compact, factual-prediction-only theory of unit-belief-conditioned local linearity. The internally frozen extension broadens the earlier observation across this fixed regression suite: all nine new datasets meet the dataset-level Cauchy-versus-Gaussian criterion under one synthetic gross-outlier regime, whereas label shuffling remains mixed. This outcome warrants severity curves, loss-only ablations, grouped or deduplicated splits, and larger-sample evaluation. It does not establish general label-noise robustness, clean-accuracy superiority, or the bounded score as the isolated source of the pattern.

A Notation and parameter conventions

A.1 Symbol table

Table 4: Symbols, domains, and interpretation.

Symbol	Domain or shape	Meaning
O, o	$\mathcal{O}$	Evidence and a realized evidence value. Evidence forms the unit belief; it is not automatically the predictor.
X, x	$\mathbb{R}^p$	Predictor and its realized value, supplied to the affine response view.
Y, y	$\mathbb{R}$	Scalar prediction target and its realization.
i	row index	Index of an observed sample/event record; not a unit identity.
U	$\mathbb{R}^d$	Ontology-level latent Unit Selection Variable. Its values distinguish candidate units or unit states through a response-relevant representation.
$u^\star$	$\mathbb{R}^d$	Conceptual world-side value of U for a unit; not assumed observed or identified.
$\widetilde{U}$	$\mathbb{R}^d$	Learner-side computational candidate integrated under the unit belief. It is not a redraw of physical identity.

Table 4 (continued)

Symbol	Domain or shape	Meaning
$Q_\phi(\,\mathrm{d}u \mid o)$	probability kernel	Unit belief after observing evidence o; not automatically a Bayesian or calibrated posterior.
$m_\phi(o)$	$\mathbb{R}^d$	Coordinatewise stable locations of the unit belief. At $\rho = 1$, these are Cauchy locations, not means.
$\gamma_\phi(o)$	$\mathbb{R}^d_{>0}$	Coordinatewise stable scales. At $\rho = 1$, these are not standard deviations and no variance exists.
E	scalar	Independent event noise with law $\mathrm{S}_\rho\mathrm{S}(0, 1)$.
ρ	$(0, 2]$	Symmetric stability index; $\rho = 1$ is Cauchy and $\rho = 2$ is Gaussian.
α_0	$\mathbb{R}$	Scalar response intercept. The subscript prevents confusion with the conventional alpha-stable family name.
β	$\mathbb{R}^p$	Population/shared slope component.
a	$\mathbb{R}^d$	Coefficient of the unit candidate in the response intercept.
B	$\mathbb{R}^{p\times d}$	Bilinear input–unit interaction matrix.
$w(u)$	$\mathbb{R}^p$	Unit-conditioned slope $\beta + Bu$.
$b(u)$	$\mathbb{R}$	Unit-conditioned intercept $\alpha_0 + a^\top u$.
$g(x)$	$\mathbb{R}^d$	Candidate coefficient $a + B^\top x$.
σ	$\mathbb{R}_{>0}$	Stable scale multiplying event noise. At $\rho = 2$, the corresponding Gaussian event standard deviation is $\sqrt{2}\sigma$.
$m(o, x)$	$\mathbb{R}$	Marginal predictive location. For Cauchy prediction it is not a conditional mean.
$s_\rho(o, x)$	$\mathbb{R}_{>0}$	Marginal stable scale. At $\rho = 2$, predictive standard deviation is $\sqrt{2}s_2(o, x)$.
$\mathcal{C}_r(o)$	subset of $\mathbb{R}^d$	Product central-quantile region used to make the unit-conditioned local view explicit.

A.2 Stable-law convention

All stable laws use the symmetric characteristic function

$$\varphi_Z(t) = \exp\{imt - |\gamma t|^\rho\}. \tag{32}$$

This convention is continuous at $\rho = 1$ because skewness is fixed at zero. The two elementary endpoints are

ρ	Stable notation	Familiar distribution
1	$\mathrm{S}_1\mathrm{S}(m, \gamma)$	$\mathrm{Cauchy}(m, \gamma)$
2	$\mathrm{S}_2\mathrm{S}(m, \gamma)$	$\mathcal{N}(m, 2\gamma^2)$

The $\rho = 2$ stable scale is therefore not the Gaussian standard deviation. Conversely, a conventional Gaussian $\mathcal{N}(m, \tau^2)$ equals $\mathrm{S}_2\mathrm{S}(m, \tau/\sqrt{2})$.

At $\rho = 1$, m and γ are location and scale. The statements "Cauchy mean m," "Cauchy variance γ^2," and "Cauchy standard deviation γ" are all false. At $1 < \rho < 2$ a first moment exists but the variance does not; at $0 < \rho \leq 1$ even the ordinary first moment does not exist. The manuscript avoids moment language and uses locations, scales, CDFs, and quantiles throughout.

B Proof of the centered local-bilinear theorem

Proof of Theorem 3. Write $z = (x, u)$, $z_0 = (x_0, u_0)$, and $\Delta = (h, k)$. The multivariate Taylor formula with Lagrange/integral remainder gives

$$f(z_0 + \Delta) = f(z_0) + Df(z_0)[\Delta] + \frac{1}{2}D^2 f(z_0)[\Delta, \Delta] + R_3, \tag{33}$$

where

$$|R_3| \le \frac{1}{6} \sup_{0 \le t \le 1} \|D^3 f(z_0 + t\Delta)\|_{\mathrm{op}} \|\Delta\|_2^3. \tag{34}$$

Partitioning the first and second derivatives into x and u blocks yields

$$Df(z_0)[\Delta] = \nabla_x f_0^\top h + \nabla_u f_0^\top k, \tag{35}$$

$$\frac{1}{2}D^2 f(z_0)[\Delta, \Delta] = \frac{1}{2}h^\top H_{xx} h + h^\top H_{xu} k + \frac{1}{2}k^\top H_{uu} k. \tag{36}$$

Subtracting Eq. (11) gives Eqs. (12) and (13). Applying $|v^\top H v| \le \|H\|_{\mathrm{op}} \|v\|_2^2$ to the pure-curvature terms gives Eq. (14).

It remains to convert the centered polynomial to the manuscript's uncentered form. Set $B = H_{xu}$ and expand:

$$T_{\mathrm{BL}}(x, u) = f_0 + f_x^\top (x - x_0) + f_u^\top (u - u_0) + (x - x_0)^\top B (u - u_0) \tag{37}$$

$$= \alpha_0 + \beta^\top x + a^\top u + x^\top B u, \tag{38}$$

where

$$\beta = f_x - B u_0, \tag{39}$$

$$a = f_u - B^\top x_0, \tag{40}$$

$$\alpha_0 = f_0 - f_x^\top x_0 - f_u^\top u_0 + x_0^\top B u_0. \tag{41}$$

The coefficients of the linearly independent monomials 1, x_k, u_j, and $x_k u_j$ are unique on any product domain with nonempty interior. This proves the final statement. □

Remark 11 (What the bound does not say). Small $\|(h, k)\|$ controls the third-order term, but the omitted pure-curvature terms are second order. The bilinear patch is therefore especially appropriate when mixed curvature is scientifically or predictively important and pure curvatures are small on the declared patch. Smoothness alone does not make the error uniformly small on a large domain.

C Stable propagation and endpoint calculations

C.1 Proof of the stable propagation theorem

Proof of Theorem 4. Fix (o, x) and abbreviate $g = g(x)$, $m_j = m_{\phi,j}(o)$, and $\gamma_j = \gamma_{\phi,j}(o)$. Conditional independence and Eq. (4) give

$$\mathbb{E}\Big[e^{itY} \mid o, x\Big] = \exp\{it(\alpha_0 + \beta^\top x)\} \prod_{j=1}^{d} \mathbb{E}\Big[e^{itg_j \widetilde{U}_j} \mid o\Big]\, \mathbb{E}[e^{it\sigma E}] \tag{42}$$

$$= \exp\{it(\alpha_0 + \beta^\top x)\} \prod_{j=1}^{d} \exp\{itg_j m_j - |t|^\rho |g_j|^\rho \gamma_j^\rho\} \exp\{-|t|^\rho \sigma^\rho\} \tag{43}$$

$$= \exp\left[it\{\alpha_0 + \beta^\top x + g^\top m_\phi(o)\} - |t|^\rho \left\{ \sigma^\rho + \sum_{j=1}^{d} |g_j|^\rho \gamma_j^\rho \right\} \right]. \tag{44}$$

This is the characteristic function of $\mathrm{S}_\rho\mathrm{S}\{m(o,x), s_\rho(o,x)\}$ under Eq. (4). Characteristic functions uniquely determine probability laws, completing the proof. □

The independence assumption is substantive. With a general multivariate symmetric stable candidate, a linear projection remains stable, but its scale is determined by the spectral measure rather than the coordinatewise sum in Eq. (17). The factorized result must not be reused for a dependent Cauchy belief without that additional model.

C.2 Cauchy density, quantiles, and bounded scores

The Cauchy(m, s) density and CDF are

$$p(y; m, s) = \frac{1}{\pi s\{1 + ((y-m)/s)^2\}}, \tag{45}$$

$$F(y; m, s) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{y-m}{s}\right). \tag{46}$$

Inverting the CDF gives Eq. (21). A central $(1-\delta)$ predictive interval is

$$\left[m - s\cot\left(\frac{\pi\delta}{2}\right), m + s\cot\left(\frac{\pi\delta}{2}\right) \right]. \tag{47}$$

For $r = y - m$, direct differentiation of Eq. (20) yields

$$\frac{\partial \ell_{\mathrm{C}}}{\partial m} = -\frac{2r}{s^2 + r^2}, \tag{48}$$

$$\frac{\partial \ell_{\mathrm{C}}}{\partial \log s} = \frac{s^2 - r^2}{s^2 + r^2}. \tag{49}$$

The function $2|r|/(s^2 + r^2)$ is maximized at $|r| = s$ with value $1/s$, and the absolute value of Eq. (49) is at most one. This proves Eq. (22). A numerical scale floor $s \geq s_{\min} > 0$ makes the location-score bound uniform.

The score with respect to the raw scale is

$$\frac{\partial \ell_{\mathrm{C}}}{\partial s} = \frac{s^2 - r^2}{s(s^2 + r^2)}, \tag{50}$$

whose bound necessarily depends on $1/s$. This is why optimization through a positive log-scale or softplus parameterization is preferable.

C.3 Gaussian endpoint

At $\rho = 2$, the characteristic function is

$$\exp\{imt - s^2t^2\} = \exp\{imt - \tfrac{1}{2}(2s^2)t^2\}, \tag{51}$$

which is the characteristic function of $\mathcal{N}(m, 2s^2)$. Its density is

$$p(y; m, s) = \frac{1}{2\sqrt{\pi}s} \exp\left\{-\frac{(y-m)^2}{4s^2}\right\}, \tag{52}$$

and taking its negative logarithm gives Eq. (25). For $r = y - m$,

$$\frac{\partial \ell_{\mathrm{G}}}{\partial m} = -\frac{r}{2s^2}, \tag{53}$$

$$\frac{\partial \ell_{\mathrm{G}}}{\partial \log s} = 1 - \frac{r^2}{2s^2}. \tag{54}$$

Unlike the Cauchy scores, these derivatives are unbounded as $|r| \to \infty$. This comparison is analytic and does not by itself establish an empirical robustness ranking.

D Learning proofs and identifiability

D.1 Fisher consistency

Proof of Theorem 7. Let P_{OX} denote the marginal law of (O, X). Subtract the risk at the true conditional density and condition on (O, X):

$$\mathcal{Q}(r) - \mathcal{Q}(r_0) = \int\int r_0(y \mid o, x) \log \frac{r_0(y \mid o, x)}{r(y \mid o, x)} \, \mathrm{d}y \, P_{OX}(\, \mathrm{d}o, \, \mathrm{d}x) \tag{55}$$

$$= \mathbb{E}_{O,X}[D_{\mathrm{KL}}\{r_0(\cdot \mid O, X) \| r(\cdot \mid O, X)\}] \geq 0. \tag{56}$$

Under the stated domination and integrability conditions, equality holds if and only if the conditional densities agree almost everywhere. The argument involves only the marginal response density; it does not constrain which latent factorization produces that density. □

If the true conditional response is Cauchy and the candidate class contains the true location and positive scale functions, identifiability of the Cauchy location–scale family makes those two response parameters Fisher consistent. It does not imply Fisher consistency of m_ϕ, γ_ϕ, a, B, and σ separately.

D.2 Varying-coefficient reduction

Proof of Proposition 8. Using $g(x) = a + B^\top x$ in Eq. (16),

$$m(o, x) = \alpha_0 + \beta^\top x + \{a + B^\top x\}^\top m_\phi(o) \tag{57}$$

$$= \{\alpha_0 + a^\top m_\phi(o)\} + x^\top \{\beta + B m_\phi(o)\} \tag{58}$$

$$= c(o) + v(o)^\top x. \tag{59}$$

Moreover, $v(o) - \beta = B m_\phi(o)$ belongs to $\mathrm{col}(B)$. The non-uniqueness follows constructively from Proposition 9. □

If the conditional response location is identified on an x-domain with nonempty interior, then $v(o)$ is its gradient with respect to x and is an observable response-law functional. Consequently,

$$\mathcal{S} = \operatorname{span}\{v(o) - v(o') : o, o' \in \mathcal{O}\} \tag{60}$$

is identifiable when those slopes are. Every UALBM factorization satisfies $\mathcal{S} \subseteq \operatorname{col}(B)$. In a minimal-rank factorization with $\operatorname{rank}(B) = \dim \mathcal{S}$, equality holds, while the basis B and coordinates $m_\phi(o)$ remain non-unique.

D.3 Latent affine reparameterization

For the next proof, a *monomial matrix* is a permutation matrix times an invertible diagonal matrix. Such transformations preserve coordinate independence up to permutation and rescaling.

Proof of Proposition 9. Let D be an invertible monomial matrix and define $V = D\widetilde{U} + c_U$. If the original coordinatewise stable belief has location $m_\phi(o)$ and scale vector $\gamma_\phi(o)$, its pushforward has location

$$m'_\phi(o) = Dm_\phi(o) + c_U \tag{61}$$

and coordinate scales obtained by applying the permutation in D and the absolute diagonal rescalings to $\gamma_\phi(o)$. It remains a product of independent $\mathrm{S}_\rho\mathrm{S}$ laws.

Set

$$a' = D^{-\top}a, \qquad B' = BD^{-1}, \tag{62}$$

$$\alpha'_0 = \alpha_0 - a^\top D^{-1}c_U, \qquad \beta' = \beta - BD^{-1}c_U. \tag{63}$$

Then $g'(x) = a' + B'^\top x = D^{-\top}g(x)$, and hence

$$\alpha'_0 + \beta'^\top x + g'(x)^\top V = \alpha_0 - a^\top D^{-1}c_U + x^\top\{\beta - BD^{-1}c_U\} + g(x)^\top D^{-1}(D\widetilde{U} + c_U) \tag{64}$$

$$= \alpha_0 + \beta^\top x + g(x)^\top \widetilde{U}. \tag{65}$$

The candidate-conditioned response is equal pointwise after transformation, and event noise is unchanged. Therefore the integrated predictive law is identical for every (o, x). □

The proposition establishes at least a signed-permutation, translation, and coordinate-rescaling ambiguity inside the factorized stable model class. A more general invertible linear transformation preserves Gaussianity if a full covariance is allowed, but does not generally preserve independent non-Gaussian stable coordinates; the proposition deliberately does not claim that larger invariance for the factorized Cauchy model.

D.4 Unit/event scale split

Proof of Proposition 10. Cauchy stability gives

$$Y \mid o \sim \text{Cauchy}\{m_\phi(o), \gamma_\phi(o) + \sigma\}. \tag{66}$$

For $0 < \delta < \sigma$, the alternative scales are positive and satisfy

$$\{\gamma_\phi(o) + \delta\} + (\sigma - \delta) = \gamma_\phi(o) + \sigma \tag{67}$$

for every o. The conditional response location and total scale are unchanged, so the laws are identical. □

Variation in $g(x)$, repeated observations, known event noise, or other restrictions can reduce this ambiguity in larger models. The proposition is a constructive counterexample to automatic identification in the unrestricted base family; it does not say that separation is impossible under every design.

E Gradients, numerical parameterization, and complexity

E.1 Forward derivatives

Write $m_U = m_\phi(o)$, $\gamma = \gamma_\phi(o)$, $g = a + B^\top x$, and

$$H = \sigma^\rho + \sum_{j=1}^{d} |g_j|^\rho \gamma_j^\rho, \qquad s = H^{1/\rho}. \tag{68}$$

The response-location derivatives are

$$\frac{\partial m}{\partial \alpha_0} = 1, \qquad \frac{\partial m}{\partial \beta} = x, \qquad \frac{\partial m}{\partial a} = m_U, \qquad \frac{\partial m}{\partial B} = x m_U^\top, \tag{69}$$

$$\frac{\partial m}{\partial m_U} = g. \tag{70}$$

Away from zero coefficients, the stable-scale derivatives are

$$\frac{\partial s}{\partial \sigma} = s^{1-\rho} \sigma^{\rho-1}, \tag{71}$$

$$\frac{\partial s}{\partial g_j} = s^{1-\rho} |g_j|^{\rho-1} \operatorname{sign}(g_j) \gamma_j^\rho, \tag{72}$$

$$\frac{\partial s}{\partial \gamma_j} = s^{1-\rho} |g_j|^\rho \gamma_j^{\rho-1}. \tag{73}$$

Chain rules through $g = a + B^\top x$ give

$$\frac{\partial s}{\partial a} = \frac{\partial s}{\partial g}, \qquad \frac{\partial s}{\partial B} = x \left(\frac{\partial s}{\partial g} \right)^\top. \tag{74}$$

At the Cauchy endpoint $\rho = 1$, $s = \sigma + \sum_j |g_j| \gamma_j$ is nondifferentiable when $g_j = 0$; any value in the valid subgradient interval $[-\gamma_j, \gamma_j]$ may be used, and the common automatic-differentiation choice zero is valid. For $0 < \rho < 1$, derivatives can be singular at zero. The implemented preliminary study trains only the Cauchy and Gaussian endpoints; the general stable theorem is a propagation result.

Combining Eqs. (69), (72) and (73) with the endpoint NLL derivatives in Appendices C.2 and C.3 gives all gradients used by ordinary reverse-mode differentiation.

E.2 Positive scales and stable computation

A practical parameterization is

$$\gamma_j(o) = \gamma_{\min} + \operatorname{softplus}\{h_{\phi,j}(o)\}, \qquad \sigma = \sigma_{\min} + \operatorname{softplus}(\eta_\sigma), \tag{75}$$

with fixed positive floors. The implementation evaluates Cauchy NLL as $\log \pi + 2 \log\{\operatorname{hypot}(r, s)\} - \log s$, avoiding explicit squaring of a very large standardized residual. Gaussian code must either use stable scale s_2 consistently or convert once to standard deviation $\sqrt{2} s_2$; mixing conventions silently changes the likelihood.

E.3 Complexity

Let $C_\phi(o)$ be the encoder cost. For one row, computing $B^\top x$ costs $O(pd)$, and locations and stable scales cost $O(d)$. Thus analytic prediction costs

$$O\{C_\phi(o) + pd + d\} \tag{76}$$

per row, or $O\{n(pd + d)\}$ beyond encoder evaluation for n rows. The model stores $pd + p + d + 2$ non-encoder scalar parameters when σ is learned. Marginalization requires no Monte Carlo samples and no matrix solve. Batching only vectorizes independent row computations; it does not introduce a support set, neighbor bank, or local refit.

F Preliminary experimental protocol and diagnostics

F.1 Study chronology and evidence partition

The experiments are preliminary engineering diagnostics rather than a tuned benchmark. The v0.2 screen on California Housing [23], Wine Quality [5], and Abalone [21] is the prior discovery group. Before running models on the v0.3 extension, we internally froze nine new datasets, seeds 42–46, corruptions, model budgets, and a dataset-level progression rule. This was an internal temporal lock, not an external preregistration. The three discovery datasets were rerun in the same artifact for comparability but were excluded from the 9-dataset gate.

For every dataset and seed, the loader uses at most 500 observations, applies a 60/20/20 train/validation/test split, and fits feature preprocessing on training rows only. Diabetes, Yacht, and Auto MPG have only 442, 308, and 392 usable rows; their split sizes are 265/88/89, 184/62/62, and 235/78/79. Other datasets contribute 300/100/100 rows. Original row identifiers are retained for split and worst-case auditing.

The paired contamination screen uses seeds 42–46 and contains

$$12 \text{ datasets} \times 5 \text{ seeds} \times 3 \text{ arms} \times 4 \text{ methods} = 720 \tag{77}$$

separately trained result rows. The arms are clean, a 20% value-changing label shuffle, and a 10% additive gross-outlier corruption. Test inputs and labels are unchanged in every arm.

F.2 Dataset provenance and parsing

Table 5 records the scalar targets and the arrays seen by the model. The UCI entries are CC BY 4.0 according to their original repository records. Diabetes is loaded through scikit-learn without extending that license statement. Concrete and Energy bytes come from OpenML mirrors of the cited UCI datasets; the frozen result mislabeled those mirrors as direct UCI raw files, and this paragraph is the append-only provenance correction. Every remote file is SHA-256 pinned in the experiment registry.

Energy uses the ninth source column as heating-load target and excludes the tenth cooling-load column from both target and predictors. Protein uses the first `RMSD` column as target and `F1`–`F9` as predictors. Auto MPG drops exactly six rows with missing horsepower, excludes car name, and one-hot encodes origin. Superconductivity reads only `train.csv`: its last column is `critical_temp`; the companion `unique_m.csv` is not supplied to the model.

F.3 Observed-interface restriction: $O = X$

The implementation retains separate predictor and evidence arguments, but in these diagnostics supplies the same observed feature vector to both: $O = X$ numerically. The evidence path produces $(m_\phi(O), \gamma_\phi(O))$, whereas the predictor path enters $g(X) = a + B^\top X$ and the affine response equation. Consequently, this study tests only the collapsed ordinary-supervision interface. It does not test distinct evidence, repeated-unit evidence, unit recovery, or generalization to other O/X relationships.

This restriction also creates an inspectable extrapolation risk. When $O = X$, the response location contains

$$\{a + B^\top x\}^\top m_\phi(x), \tag{78}$$

so an out-of-range predictor can be amplified through both paths. This is a property of the fitted response map, not latent retrieval or an inference-time local solve.

F.4 Predictors and optimization

All neural methods use a two-layer SiLU perception network of width 64 and an eight-dimensional linear hidden or unit-location head. Direct MLP concatenates its deterministic hidden representation with X, applies a linear scalar output head, and minimizes MSE. It defines neither a unit belief nor a predictive stable law.

Gaussian and Cauchy UALBM have identical encoder and head dimensions. Each has an eight-dimensional location head, a positive stable-scale head, and the same bilinear response parameterization. Unit and event scales use softplus with a floor of 10^{-4}; their initial standardized scales are 0.2 and 0.1. The Gaussian endpoint minimizes its matched Gaussian NLL with conventional standard deviation $\sqrt{2}s_2$. The Cauchy endpoint minimizes Cauchy NLL with scale s_1. Endpoint weights are learned independently, although runs use the same seed and initialization procedure.

The contextual tree reference is `sklearn.ensemble.GradientBoostingRegressor` [25] with library defaults and `random_state` equal to the dataset seed. It is not tuned and is not a topology-matched ablation. It is fitted on the training split only; validation labels do not select a tree checkpoint.

Neural optimization is implemented in PyTorch [24] and uses Adam with learning rate 10^{-3}, zero weight decay, batches of 128, at most 80 epochs, and early stopping with patience 12 and minimum improvement 10^{-6}. Each neural arm fits its target location and population standard deviation from that arm's training labels. Dirty arms therefore change both the fitted target transform and the validation checkpoint. This evaluates the complete dirty-development pipeline rather than isolating the endpoint loss.

F.5 Auditable label corruptions

Corruptions are generated after splitting. Train and validation use independent deterministic random streams indexed by dataset, seed, arm, and split. Every method receives the same frozen corrupted labels for a given dataset–seed–arm cell.

For the shuffle arm, exactly 20% of train and validation labels are selected. A value-changing permutation is constructed so that every selected numerical label differs from its original value, including for repeated integer-valued targets.

For the gross-outlier arm, exactly 10% of train and validation labels are selected and replaced by

$$y_i^{\text{dirty}} = y_i + \epsilon_i \, 5\widehat{\sigma}^{\text{clean}}_{y,\text{train}}, \qquad \epsilon_i \in \{-1, +1\}. \tag{79}$$

The signs are count-balanced, the validation mask and signs are independent of the training plan, and both splits use the clean training-label population standard deviation as reference. Hash checks

verify the selected row IDs, corrupted labels, applied perturbations, shared labels across methods, and unchanged test labels.

F.6 Paired endpoints

For method m, dataset d, seed s, and dirty arm c, define

$$\Delta_{m,d,s,c} = \mathrm{RMSE}^{c}_{m,d,s} - \mathrm{RMSE}^{\mathrm{clean}}_{m,d,s}\,. \tag{80}$$

The endpoint-matched contrast is

$$D_{d,s,c} = \Delta_{\mathrm{C},d,s,c} - \Delta_{\mathrm{G},d,s,c}. \tag{81}$$

Negative D means that Cauchy lost less clean-test accuracy than Gaussian for that paired seed. It is always accompanied by the absolute dirty-arm gap $\mathrm{RMSE}^{c}_{\mathrm{C}} - \mathrm{RMSE}^{c}_{\mathrm{G}}$; a smaller degradation against an unstable clean anchor is not interpreted as beneficial corruption.

Because the study has only five seeds and contains one high-leverage test point, medians, raw paired seeds, and win counts are primary. Means are not used alone. RMSE, MAE, median, 90th-percentile and maximum absolute error, prediction range, and the original ID of the worst test row are retained. No null-hypothesis or population-level robustness inference is made.

F.7 Dataset-level extension gate and result

For dataset d and corruption c, define the two within-dataset medians

$$T^{\mathrm{deg}}_{d,c} = \mathrm{median}^{46}_{s=42}\,\frac{D_{d,s,c}}{s_{y,d,s}}, \tag{82}$$

$$T^{\mathrm{dirty}}_{d,c} = \mathrm{median}^{46}_{s=42}\,\frac{\mathrm{RMSE}^{c}_{\mathrm{C},d,s} - \mathrm{RMSE}^{c}_{\mathrm{G},d,s}}{s_{y,d,s}}, \tag{83}$$

where $s_{y,d,s}$ is the clean-training target population standard deviation. A prospective dataset is supportive only if both quantities are negative. The frozen gross-outlier rule called 8–9 of 9 supportive datasets a strong suite-specific signal worth severity and ablation work, 5–7 heterogeneous, and 0–4 no broadened signal. The primary aggregation units are the nine dataset medians, not 45 independent experiments.

The gross-outlier arm passes the internal progression rule; the shuffle arm does not broaden the observation. MAE agrees on both directions for 9/9 gross datasets. Secondary tail summaries in Table 7 show that the gross direction is present through most of the error distribution, whereas shuffle directions split. These are descriptive sensitivities rather than additional frozen decision criteria.

After seeing results, we computed a 10,000-repetition hierarchical bootstrap that resamples datasets and then seeds, and leave-one-dataset-out macro medians. For the gross arm, the descriptive 95% bootstrap intervals are $[-0.267023, -0.082202]$ for degradation contrast and $[-0.296489, -0.079770]$ for dirty gap; every leave-one-dataset-out macro contrast is negative. Both shuffle intervals cross zero. These post-result summaries are not a confirmatory significance analysis.

F.8 Random-row duplicate-feature limitation

Row identifiers never cross partitions, but identical predictor vectors can. Across seeds 42–46, train–test intersections contain 12 distinct vectors in Concrete, 59 in QSAR Fish, one in Protein, and

Table 5: Dataset provenance. "Full" is the parsed usable row count and "used" is the per-seed cap. Predictors are the columns after documented filtering or encoding.

Dataset	Full	Used	p	Target / preprocessing
California [23]	20,640	500	8	median house value
Wine [5]	6,497	500	11	quality; red and white combined
Abalone [21]	4,177	500	10	rings; sex one-hot encoded
Diabetes [6]	442	442	10	one-year disease progression
Concrete [31]	1,030	500	8	compressive strength
Energy [30]	768	500	8	heating load; cooling load excluded
Airfoil [3]	1,503	500	5	scaled sound-pressure level
Yacht [10]	308	308	6	residuary resistance
Auto MPG [26]	392	392	9	MPG; 6 missing-horsepower rows dropped
QSAR Fish [1]	908	500	6	LC50
Protein [27]	45,730	500	9	RMSD, read from the first column
Superconductivity [13]	21,263	500	81	critical temperature

Table 6: Complete prospective endpoint comparison. Values are five-seed medians in clean-training target-SD units.

	Gross outlier		Shuffle		Seeds satisfying both	
Dataset	T^{deg}	T^{dirty}	T^{deg}	T^{dirty}	Gross	Shuffle
Diabetes	-0.157925	-0.175861	-0.004035	0.006517	5/5	2/5
Concrete	-0.160053	-0.247693	-0.004218	-0.080971	4/5	4/5
Energy (heating)	-0.096795	-0.090246	-0.006161	-0.010887	4/5	2/5
Airfoil	-0.170705	-0.223965	0.003490	-0.049322	5/5	2/5
Yacht	-0.319310	-0.463364	0.015871	-0.180775	5/5	2/5
Auto MPG	-0.125786	-0.114303	0.001716	-0.019306	5/5	1/5
QSAR Fish	-0.230646	-0.218305	-0.011010	-0.000825	4/5	3/5
Protein	-0.072127	-0.049454	-0.002240	0.055109	4/5	0/5
Superconductivity	-0.267023	-0.296489	-0.001305	-0.002665	5/5	3/5
Supportive datasets	9/9		4/9		41/45	19/45
Macro dataset median	-0.160053	-0.218305	-0.002240	-0.010887		

Table 7: Prospective tail sensitivity. Each count is out of nine datasets after taking the within-dataset median over five seeds. "Both" means that Cauchy has both a lower dirty metric and lower degradation than Gaussian.

Arm	Metric	Lower dirty	Lower degradation	Both
Gross outlier	Median absolute error	9	9	9
Gross outlier	90th-percentile absolute error	8	9	8
Gross outlier	Maximum absolute error	9	9	9
Shuffle	Median absolute error	8	4	4
Shuffle	90th-percentile absolute error	7	6	6
Shuffle	Maximum absolute error	4	7	3

27 in Superconductivity. Most QSAR Fish and Superconductivity overlaps have different targets, so this is not simple duplicate-(X, Y) target leakage; it does mean that the experiment is a random-row contamination screen rather than strict unseen-X or unseen-unit generalization. Diabetes, Energy, Airfoil, Yacht, and Auto MPG have no such cross-split exact-X matches. The next gate will group or deduplicate exact predictor vectors; Superconductivity can additionally use material identities from `unique_m.csv`. Protein-level grouping requires an external identity mapping absent from the current CASP file.

F.9 Implementation and numerical diagnostics

A float64 oracle checks a hand-computed bilinear example, analytic stable-scale propagation, automatic gradients, singleton/batch equivalence, and independent latent simulations in which unit coordinates and event noise are sampled separately. The maximum observed gradient discrepancy was 1.22×10^{-15}; with 50,000 draws, the Gaussian and Cauchy probability-integral-transform Kolmogorov–Smirnov distances were 0.00451 and 0.00388.

Cauchy NLL is evaluated as

$$\log \pi + 2 \log\{\mathrm{hypot}(y - m, s)\} - \log s, \tag{84}$$

which avoids explicitly squaring a very large residual. Gaussian likelihood uses $\sqrt{2}s_2$ consistently, and its residual-square arithmetic is promoted to float64 before division so a float32 intermediate cannot create artificial overflow. Training aborts on a non-finite loss, and all reported outputs were finite. No prospective run overflowed. In the extension, 31/45 clean and 27/45 shuffle Gaussian runs selected the final epoch of the fixed 80-epoch budget, compared with 8/45 and 12/45 Cauchy runs. This records fixed-budget behavior, not convergence or a theorem that Cauchy optimization is smoother.

The tail diagnostics nevertheless exposed a response-extrapolation failure. For California seed 45, original test row 9172 has standardized average occupancy 113.407, while the largest absolute training value for that feature is 4.457. On the clean arm, Gaussian UALBM predicts -201.509 for target 1.546, and Cauchy UALBM predicts -17.881. The Gaussian absolute error is 203.055 although its MAE and 90th-percentile absolute error are 2.525 and 1.000. This is finite bilinear $O = X$ extrapolation, not floating-point overflow, and motivates reporting tail diagnostics in addition to RMSE.

F.10 Reproducibility and claim limitations

The recorded run used Python 3.12.2, PyTorch 2.13.0, scikit-learn 1.7.2, NumPy 2.2.6, SciPy 1.15.3, and CPU execution. The frozen seeds, hyperparameters, row-ID hashes, corruption-plan hashes, raw

per-seed results, and analysis artifacts are retained with the experimental package. The raw 720-row v0.3 artifact has SHA-256 abbreviated `5e607fcd...85e585`; its internally frozen configuration has SHA-256 abbreviated `5a9735e6...798126`. The evidence manifest records both full digests. The raw artifact keeps its original `diagnostic_only` and "not manuscript evidence" fields as historical provenance; a later owner promotion note admits only the bounded result reported here without rewriting those fields. The original artifact did not hash every P32 parser and training source file, so exact source-manifest capture is required for the next run.

These experiments do not establish universal predictive superiority, general label-noise robustness, calibrated uncertainty, or recovery of a true unit. The fixed-default tree and Direct MLP are contextual baselines; only Gaussian and Cauchy UALBM form a topology-matched endpoint comparison. The dirty target transform and dirty validation selection prevent attributing the observed outlier pattern solely to the bounded Cauchy score. Learning curves and gradient norms were not retained, so the study cannot establish smoother Cauchy optimization. Only one symmetric gross-outlier severity was tested; there is no natural, asymmetric, heteroscedastic, or adversarial noise study. Wine and Abalone NLLs are continuous working-density scores for integer-valued targets, not proper discrete likelihoods. The study also does not test $O \neq X$ or any causal or counterfactual interpretation.

G Reader FAQ

G.1 What exactly is a unit belief?

The Unit Selection Variable U first specifies the model's unit ontology: its values answer which candidate unit or unit state is under consideration and encode its response-relevant representation. A sample is an observed event/record that carries evidence about this object; its row index is not the unit identity. The unit belief is the learner-side probability kernel $Q_\phi(\,\mathrm{d}u \mid O = o)$ over candidate values of U after seeing evidence. A point embedding is recovered as a degenerate special case, but a nondegenerate belief also carries scale into the predictive law. Without a prior, an evidence likelihood, and calibration evidence, the term "Bayesian posterior" would be stronger than the model establishes.

G.2 Is unit selection the same as sample selection?

No. Unit selection first asks which ontology-level value of U characterizes the unit associated with an observed event. Evidence in row (o_i, x_i, y_i) yields the learner-side answer $Q_\phi(\,\mathrm{d}u \mid o_i)$. A candidate u induces a local coefficient view only downstream, through $u \mapsto \{w(u), b(u)\}$; the view is not the definition of the unit. The predictive computation then marginalizes that belief:

$$\ell_i(\theta, \phi) = -\log \int p_\theta(y_i \mid x_i, u)\, Q_\phi(\,\mathrm{d}u \mid o_i). \tag{85}$$

The kernel $Q_\phi(\cdot \mid o_i)$ is normalized over unit candidates and changes that row's coefficient and response law. It does not decide whether row i appears in the data.

Sample selection instead operates across rows. If ω_i denotes a source-row importance weight, the corresponding weighted minibatch objective would be

$$\mathcal{L}_\omega = \frac{\sum_{i \in B} \omega_i\, \ell_i(\theta, \phi)}{\sum_{i \in B} \omega_i}. \tag{86}$$

The inner integral in Eq. (85) is unit-belief marginalization; the outer average in Eq. (86) is sample weighting. The current P32 baseline sets $\omega_i = 1$ and performs no sample-selection correction.

The ontology permits one unit to be associated with multiple observed events or repeated measurements. Conversely, even if a dataset happens to contain one row per unit, the row and the unit remain different kinds of objects: the former is an observation record, whereas the latter is what the record is about.

For a possible source-to-target prediction extension, let $W = (O, X)$ and suppose

$$P_{\mathrm{t}}(Y \mid W) = P_{\mathrm{s}}(Y \mid W), \qquad P_{\mathrm{t}}^{W} \ll P_{\mathrm{s}}^{W}. \tag{87}$$

Then the density ratio $r(w) = \mathrm{d}P_{\mathrm{t}}^{W}(w)/\,\mathrm{d}P_{\mathrm{s}}^{W}(w)$ can reweight the source predictive risk. It also changes the aggregate unit belief by

$$\overline{Q}_{\phi,\mathrm{t}}(A) = \frac{\mathbb{E}_{P_{\mathrm{s}}}[r(W)Q_{\phi}(A \mid O)]}{\mathbb{E}_{P_{\mathrm{s}}}[r(W)]}. \tag{88}$$

This is a change-of-measure identity under the stated overlap and conditional invariance assumptions, not a claim that Q_{ϕ} is a sample propensity, calibrated latent posterior, or valid unit-based transport weight. Directly using a ratio of aggregate unit densities as a row weight would require a stronger latent transport model and additional calibration evidence. When weights are inverse inclusion probabilities the construction may be called IPW; when $P_{\mathrm{t}}^{W}/P_{\mathrm{s}}^{W}$ is estimated directly, *importance weighting* or *density-ratio weighting* is more precise. Extreme ω_i also multiply the per-row score, so a bounded Cauchy residual score does not remove the need to inspect overlap, weight tails, effective sample size, and clipping or stabilization bias.

G.3 Does sampling a candidate redraw the unit?

No. $\widetilde{U} \sim Q_{\phi}(\cdot \mid o)$ is a computational device for representing the learner's uncertainty. In the stable model the integral is analytic, so an implementation need not sample at all.

G.4 How is UALBM fundamentally different from a VAE?

A variational autoencoder specifies a generative latent-variable model such as $p(z)p_{\theta}(O \mid z)$. Its encoder $q_{\phi}(z \mid O)$ approximates a posterior, and its evidence lower bound trades a reconstruction or data-likelihood term against a per-observation KL term to the latent prior. UALBM instead specifies a supervised factual predictor $p(Y \mid X, O)$. Its encoder outputs the unit belief $Q_{\phi}(U \mid O)$, which changes the affine response coefficients $w(U)$ and $b(U)$; stable closure then integrates that belief directly into the response location and scale. Training minimizes predictive NLL and contains no $p_{\theta}(O \mid U)$ decoder, evidence reconstruction, ELBO, or per-observation prior KL.

Both models may use an amortized encoder whose output is a probability distribution, so the diagrams can look superficially similar. The fundamental difference is where the model places structure and what likelihood it learns, not whether the latent law is Gaussian or Cauchy. A VAE could use a Cauchy latent, and UALBM has a Gaussian endpoint. Conversely, adding a prior and a generative evidence decoder to UALBM would make that extension more VAE-like. The current predictive unit belief is not thereby identified as a physical unit or certified as a Bayesian posterior. The rejected aggregate characteristic-function objective is also not an ELBO term and was not used in the reported experiments.

G.5 Why call the model abductive?

"Abductive" names the evidence-to-unit-belief step: the learner reasons from observed evidence to candidate unit descriptions before prediction. The word does not imply Bayesian posterior correctness, physical identity recovery, or causal abduction.

G.6 What does local mean?

For a fixed candidate u, $x \mapsto b(u) + w(u)^\top x$ is affine. The unit belief's location and central-quantile region define which candidate-conditioned affine views matter for evidence o. This is unit-conditioned locality. It is not an input-neighborhood estimator, and the Taylor theorem is approximation motivation rather than an operational query-time fitting rule.

G.7 Is there a top-k or neighborhood-weighted solve?

No. UALBM has no reference bank, distance function, top-k selection, prototype, local Gram matrix, weighted least-squares system, or inference-time optimization. Prediction is one encoder pass followed by the analytic formulas for $m(o, x)$ and $s_\rho(o, x)$.

G.8 How can a bilinear model represent global nonlinearity?

The encoder may map evidence nonlinearly to the location and scale of the unit belief. After marginalization, both the predictive location and scale can therefore be nonlinear functions of observed evidence. Conditional on a candidate unit, however, the x-view is affine. This is a structured inductive bias, not a theorem that every global nonlinear function is exactly representable by one finite-dimensional bilinear model.

G.9 Why is the Cauchy location not called a mean?

A nondegenerate Cauchy variable has no finite expectation or variance. Its location is its median and mode under the standard symmetric parameterization, and its scale controls quantile spread. The paper therefore reports location, scale, CDF, and quantiles.

G.10 How is the Gaussian baseline matched?

Both endpoints use $\exp\{imt - |\gamma t|^\rho\}$. At $\rho = 2$, the variance is $2\gamma^2$ and the standard deviation is $\sqrt{2}\gamma$. Matching requires the same bilinear coefficients and encoder structure plus this scale conversion; using γ directly as a Gaussian standard deviation would be mismatched.

G.11 How do the three proposed neural architectures differ?

Direct MLP is an ordinary point-prediction baseline. Evidence passes through a perception MLP and linear hidden representation head, predictor x joins at a scalar output head, and the model is trained with a point loss such as MSE. It defines no unit belief and no stable marginal law. It is therefore an algorithm-level baseline, not a one-factor ablation of UALBM.

Gaussian and Cauchy UALBM use the same architecture template: perception MLP, linear unit-location head $m_\phi(o)$, positive stable-scale head $\gamma_\phi(o)$, and the analytic bilinear marginalizer. They use independent weights, but can be matched in topology, dimension, and parameter count. Their only intended endpoint differences are the common stability index used for both U and E, the resulting scale aggregation, and the NLL. The architecture never calls Cauchy location a mean, and the Gaussian output standard deviation is $\sqrt{2}\, s_2$, not s_2.

G.12 Does closed form mean closed-form training?

No. It means the latent candidate can be integrated analytically and the Cauchy/Gaussian endpoint likelihood is elementary. Neural and bilinear parameters are learned by numerical maximum

likelihood.

G.13 How do training and inference differ?

For a minibatch $\mathcal{I}$, training computes the row-wise analytic parameters $m(o_i, x_i)$ and $s_\rho(o_i, x_i)$, aggregates $|\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \ell(y_i; m_i, s_i)$, and backpropagates only through those batch rows. It does not inspect the full training set, retrieve neighbors, or build a support/query episode. At inference, learned parameters are frozen and the same analytic forward map is evaluated for each (o, x). Batching inference rows is only vectorization; no retrieval or local solve appears in either stage.

G.14 Can evidence and predictor be the same vector?

Yes. An ordinary supervised application may set $O = X$ numerically. The roles remain distinct in the architecture: the evidence path produces the unit belief; the predictor path enters the affine response equation. Keeping these interfaces separate makes the restriction inspectable.

G.15 Is the scale calibrated uncertainty?

Not automatically. The marginal response scale is a valid model parameter and can be assessed with held-out log scores, coverage, or calibration diagnostics. Calling its unit component calibrated epistemic uncertainty requires that assessment and, because of Proposition 10, possibly additional data or constraints separating unit and event contributions.

G.16 Why is this not a counterfactual or causal paper?

X is only a predictor and Y only a factual prediction target. The paper defines no treatment, intervention, potential outcome, cross-world quantity, or identification condition. Evaluating the fitted formula at a different numerical x is model evaluation; it is not given a causal interpretation.

G.17 What has been empirically established?

Only a bounded, small-sample diagnostic result. The implementation oracle supports agreement between the code and the stated bilinear and stable-law formulas. The original three-dataset discovery screen generated a gross-outlier hypothesis. After an internal pre-result lock, all nine new datasets met the two-condition, dataset-median Cauchy-versus-Gaussian rule under one 10% balanced $\pm 5 s_y$ development-label protocol; 41/45 seed cells agreed on both directions. Under exact 20% value-changing shuffle, only 4/9 datasets and 19/45 seed cells met the same descriptive rule. On clean extension data, fixed-default gradient boosting won seven dataset medians and Direct MLP two; neither stable endpoint won. The evidence is therefore a suite-specific synthetic gross-outlier pattern, not general label-noise robustness or predictive superiority. The high-leverage California point, duplicate-feature random-split limitation, fixed epoch budget, and full-pipeline rather than loss-only comparison in Appendix F remain part of the empirical result.

G.18 When should the work be described as a technical note?

If a literature audit or future experiment shows that the unit-belief composition creates no meaningful distinction from standard varying-coefficient or distributional regression, the honest positioning is a technical note that packages classical components. The manuscript should not claim a new paradigm merely because the parameterization uses new names.

References

[1] Davide Ballabio, Viviana Consonni, and Roberto Todeschini. QSAR fish toxicity. UCI Machine Learning Repository, 2015. CC BY 4.0; https://doi.org/10.24432/C5JG7B.

[2] Léon Bottou and Vladimir Vapnik. Local learning algorithms. *Neural Computation*, 4(6): 888–900, 1992. doi: 10.1162/neco.1992.4.6.888.

[3] Thomas Brooks, D. Pope, and Michael Marcolini. Airfoil self-noise. UCI Machine Learning Repository, 1989. CC BY 4.0; https://doi.org/10.24432/C5VW2C.

[4] William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403): 596–610, 1988. doi: 10.1080/01621459.1988.10478639.

[5] Paulo Cortez, A. Cerdeira, Fernando Almeida, Telmo Matos, and J. Reis. Wine quality. UCI Machine Learning Repository, 2009. https://doi.org/10.24432/C56S3T.

[6] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

[7] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London, 1996. doi: 10.1201/9780203748725.

[8] Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179–195, 2008. doi: 10.4310/SII.2008.v1.n1.a15.

[9] Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

[10] J. Gerritsma, R. Onnink, and A. Versluis. Yacht hydrodynamics. UCI Machine Learning Repository, 1981. CC BY 4.0; https://doi.org/10.24432/C5XG7R.

[11] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

[12] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

[13] Kam Hamidieh. Superconductivity data. UCI Machine Learning Repository, 2018. CC BY 4.0; https://doi.org/10.24432/C53P47.

[14] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–796, 1993. doi: 10.1111/j.2517-6161.1993.tb01939.x.

[15] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

[16] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

[17] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkE6PjC9KX.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

[19] Nadja Klein, Thomas Kneib, Stefan Lang, and Alexander Sohn. Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, 9(2):1024–1052, 2015. doi: 10.1214/15-AOAS823.

[20] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. doi: 10.2307/2529876.

[21] Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1994. https://doi.org/10.24432/C55C7W.

[22] John P. Nolan. *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer, Cham, 2020. doi: 10.1007/978-3-030-52915-4.

[23] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019. URL https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

[26] R. Quinlan. Auto mpg. UCI Machine Learning Repository, 1993. CC BY 4.0; https://doi.org/10.24432/C5859H.

[27] Prashant Rana. Physicochemical properties of protein tertiary structure. UCI Machine Learning Repository, 2013. CC BY 4.0; https://doi.org/10.24432/C5QW3H.

[28] Robert A. Rigby and D. Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005. doi: 10.1111/j.1467-9876.2005.00510.x.

[29] Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York, 1994. doi: 10.1201/9780203738818.

[30] Athanasios Tsanas and Angeliki Xifara. Energy efficiency. UCI Machine Learning Repository, 2012. CC BY 4.0; https://doi.org/10.24432/C51307.

[31] I-Cheng Yeh. Concrete compressive strength. UCI Machine Learning Repository, 1998. CC BY 4.0; https://doi.org/10.24432/C5PK67.