Proof of Concept Paper Founding Statement

Linear Networks A Gradient-Free Architecture For AI Implicit

Momen Ghazouani *Chief Scientist, Setaleur Aplamda*

April 27, 2026

Setaleur Aplamda on GitHub

Abstract

We introduce **Linear Networks**, a gradient-free neural architecture class designed to serve as a computational substrate for the AI Implicit research paradigm a framework that measures intelligence through the extraction, compression, and transfer of implicit structure, rather than through direct input-output prediction. Where conventional architectures learn surface-level correlations by minimizing differentiable loss functions, Linear Networks build knowledge by constructing multi-space relational density fields across semantic, logical, and compositional representation spaces, then propagating parallel Relational Inference Chains to evaluate structural hypotheses through relational coherence without any gradient signal. The central architectural contribution is a four-space relational density framework: a *Local Density Matrix* encoding geometric neighborhood topology, a *Directed Cauchy Affinity Matrix* encoding asymmetric discriminative relationships, a *Spurious Feature Matrix* encoding surface-level correlations to be monitored, and a *Compositional Relational Density* encoding multi-hop structural chains. Critically, the aggregation of these spaces into an Epistemic Confidence signal Ψ operationalizes AI Implicit Principle 3 : a system should know when it does not know.

We evaluate Linear Networks on a controlled Colored MNIST benchmark designed to measure epistemic properties under distribution shift and spurious correlation. At proof-of-concept scale, and relative to baselines operating on identical features, Linear Networks demonstrate: competitive predictive accuracy (0.970 standalone, 0.978 hybrid); an Epistemic AUROC of 0.897substantially exceeding the KNN reference (0.669); Selective Prediction Quality of 0.986 versus 0.035 for KNN; calibrated Expected Calibration Error of 0.020 (from a raw 0.722 prior to post-hoc isotonic calibration); 100% out-of-distribution rejection on Gaussian noise; and Cross-Distribution Retention of 0.995 versus 0.146 for KNN under five-shot reduction. A complementary experience-efficiency benchmark reveals that Linear Networks do not outperform optimized gradient-based baselines in standard accuracy-per-sample terms, an honest limitation we document and theoretically contextualize as an expected consequence of the architecture's epistemic rather than predictive design objective. These results constitute

a proof of concept establishing that multi-space relational density estimation, without gradient descent, is a viable inductive bias for learning that targets epistemic awareness, calibrated uncertainty, and structural retention under distributional shift. This paper lays the architectural and empirical foundation for the Linear Networks research program within AI Implicit.

Keywords: Linear Networks, AI Implicit, Bold Learning, Relational Density Estimation, Epistemic Confidence, Gradient-Free Learning, Out-of-Distribution Rejection, Experience Compression

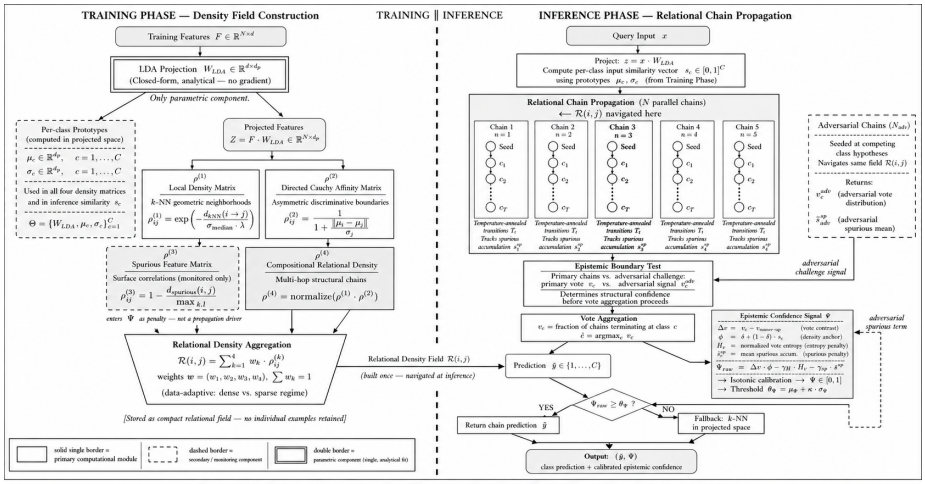

Figure 1. **Linear Networks Architecture**. **Left**: *gradient-free density field construction during training.* **Right**: *Relational Inference Chain propagation and Epistemic Confidence computation at inference time. The horizontal arrow carries the built relational field $\mathcal{R}(i,j)$ from Training into Chain Propagation — the system's single structural transfer point. $\rho^{(3)}$ (dashed) and adversarial chains enter Ψ as penalty signals, not as primary propagation drivers.*

Figure 1 : Linear Networks Architecture *Left : gradient-free density field construction during training. Right: Relational Inference Chain propagation and Epistemic Confidence computation at inference time. The horizontal arrow carries the built relational field $R(i,j)$ from training into Chain Propagation the system's single structural transfer point. $\rho^{(3)}$ (dashed) and adversarial chains enter as penalty signals, not as primary propagation drivers*

---

1. Introduction

1.1 Two Views of Learning

There are two fundamentally different conceptions of what a learning system should do.

The first, and dominant, view holds that learning is function approximation: given inputs $\mathbf{x}$ and targets $\mathbf{y}$, find parameters θ that minimize a loss $\mathcal{L}(f_\theta(\mathbf{x}), \mathbf{y})$. This view has produced systems that match or exceed human performance on image recognition, strategic game-playing, language translation, and protein

structure prediction. Modern large-scale neural networks are its highest expression.

The second view holds that learning is structure extraction: given experience, discover the compressed relational patterns that generate observed phenomena, and organize them into representations that transfer across contexts. This view is older it traces through statistical mechanics, information theory, and cognitive science but has not yet found a canonical neural architecture that realizes it without falling back on gradient-based optimization.

We argue that the second view is not merely an alternative framing of the first. It is a harder problem with different success criteria. Meeting those criteria requires: (1) a mechanism for constructing geometry-aware relational fields across multiple representation spaces without loss minimization; (2) a procedure for propagating parallel structural hypotheses through those fields; and (3) a signal that quantifies the system's confidence in its own structural assignments and that is capable of registering genuine uncertainty rather than producing forced predictions on all inputs.

This paper introduces **Linear Networks** as a proof-of-concept architectural realization of these requirements within the AI Implicit paradigm [Ghazouani, 2026a].

1.2 The AI Implicit Context

AI Implicit [Ghazouani, 2026a] is a research paradigm that operationalizes intelligence as the capacity to extract, compress, and transfer tacit knowledge. It rests on three principles:

- **Principle 1 Tacit Structure Extraction:** The primary learning signal should be the recovery of latent relational structure, not prediction accuracy on surface tokens.
- **Principle 2 Experience Compression:** Intelligence is measured by knowledge density how much reusable structure is extracted per unit of experience.
- **Principle 3 Epistemic Confidence:** A system that cannot recognize the boundaries of its own knowledge is not intelligent; it is brittle. Genuine learning requires uncertainty awareness.

Linear Networks are situated within this paradigm and embody the Bold Learning philosophy [Ghazouani, 2026b]: the commitment to infer latent structural hypotheses beyond observed surface patterns. Unlike gradient-based realizations of Bold Learning such as Deep Transducers [Ghazouani, 2026], which refine structural prototypes through iterative reconstruction pressure Linear Networks realize this commitment through static relational density estimation, eliminating gradient dependence entirely. This is a deliberate architectural choice, not a limitation. Deep Transducers learn to compress structure by training prototype geometry over many iterations; Linear Networks commit to a structural organi-

zation in a single estimation pass over the training distribution, then navigate that organization at inference time through Relational Inference Chains. Both share the AI Implicit goal of compressing experience into reusable structure. They differ in mechanism and in the nature of their structural commitment.

1.3 The Core Argument

The thesis of this paper is precise: relational density estimation across multiple geometrically distinct spaces, without gradient descent, is a sufficient inductive bias for building learning systems with calibrated epistemic awareness and robust structural retention under distributional shift. This claim has two parts. The positive part that relational density estimation produces meaningful epistemic signals is supported by the empirical results of Section 6. The negative part that standard optimization-based systems do not provide these properties by design is a consequence of the optimization framework itself. Any system trained by minimizing expected loss on a fixed distribution will assign confidence according to learned decision boundaries, not according to the structural distance between a new input and the known training distribution. The result is that optimization-based systems are, in the language of the AI Implicit paradigm, epistemically opaque: they cannot, without substantial architectural modification, register the difference between a familiar input and an unfamiliar one.

Linear Networks address this by constructing the knowledge representation as an explicit relational density field, where the Epistemic Confidence signal Ψ is derived directly from the density of the inference chains that converge on a prediction. When no chains converge with sufficient density because the input is genuinely unfamiliar Ψ is low by construction. This is not a post-hoc calibration trick; it is an architectural property of the density estimation procedure.

1.4 Contributions

This paper makes the following contributions:

1. We introduce **Linear Networks** as a gradient-free neural architecture class for structure-centric learning, grounding the design in AI Implicit principles.
2. We formalize a **four-space relational density framework** Local Density Matrix, Directed Cauchy Affinity Matrix, Spurious Feature Matrix, and Compositional Relational Density as the core representational mechanism.
3. We introduce an **Epistemic Confidence signal** Ψ that produces a calibrated uncertainty estimate from the geometry of relational chain convergence, in a single inference pass.
4. We define and evaluate the **Experience-Compressed Intelligence (ECI)** metric suite Epistemic AUROC, Selective Prediction Quality (SPQ), Cross-Distribution Retention (CDR), and calibration quality as

the appropriate evaluation framework for Linear Networks.
5. We provide honest empirical evaluation on two benchmarks, documenting both the demonstrated epistemic properties and the predictive limitations of the current proof of concept.

2. Background and Related Work

2.1 Gradient-Based Learning and Its Epistemic Limitations

The dominant paradigm in contemporary machine learning rests on the principle of empirical risk minimization:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\mathcal{L}(f_\theta(\mathbf{x}), \mathbf{y})\right]$$

Through stochastic gradient descent and its variants, this framework has produced systems of extraordinary capability. However, the same framework produces a systematic epistemic blind spot: when a learned model f_{θ^*} encounters an input that lies far outside its training distribution, it produces a prediction through the same softmax-normalized output mechanism it uses for familiar inputs. The resulting confidence values reflect the geometry of learned decision boundaries, not the structural novelty of the input. This is the epistemic opacity problem [Ghazouani, 2026] confident predictions on unfamiliar inputs are not a bug but a structural consequence of the training objective. Several approaches have been proposed to address this limitation post-hoc. Bayesian neural networks [Blundell et al., 2015] place distributions over weights, enabling principled uncertainty propagation at substantial computational cost. Monte Carlo Dropout [Gal and Ghahramani, 2016] approximates posterior inference by treating stochastic forward passes as variational inference. Deep ensembles [Lakshminarayanan et al., 2017] estimate uncertainty through disagreement among independently trained models. Mahalanobis distance methods [Lee et al., 2018] detect out-of-distribution inputs through statistical distance in activation space.

These methods are powerful, but they address the epistemic opacity problem at the cost of either computational overhead (multiple forward passes or multiple models) or post-hoc calibration on held-out data. They do not change the fundamental architecture; they wrap it with additional estimation procedures. Linear Networks take a different approach : epistemic awareness is an architectural property of the density estimation procedure, not a post-hoc addition. The Epistemic Confidence signal Ψ is derived from the same relational density fields used for prediction, in a single inference pass.

2.2 Prototype-Based and Non-Parametric Learning

The use of learned prototypes as representational anchors has a long history. Prototype networks [Snell et al., 2017] demonstrate that class means in embedding space are powerful representational structures for few-shot generalization.

k-nearest neighbor classifiers provide a non-parametric baseline that uses training examples directly at inference time. Both approaches represent knowledge through structural relationships in feature space rather than through learned decision boundaries, placing them closer in spirit to Linear Networks than standard parametric models.

However, both standard prototypes and k-nearest neighbor classifiers operate in a single geometric space, and neither provides a calibrated epistemic signal. The k-nearest neighbor margin the fraction of neighbors supporting the winning class is a coarse confidence proxy that collapses to zero discriminative power in the few-shot regime (as confirmed by our experiments, where KNN achieves 0.035 SPQ versus 0.986 for Linear Networks). Linear Networks extend the prototype intuition to multiple relational spaces simultaneously and derive epistemic confidence from the convergence structure of inference chains rather than from vote margins.

2.3 Relational Reasoning and Graph-Based Methods

Graph neural networks [Kipf and Welling, 2017; Xu et al., 2019] operate on relational data through message passing, offering a powerful framework for structured reasoning. However, GNNs are trained through gradient-based optimization over fixed graph structures. The relational structure must be specified prior to training and remains fixed. Linear Networks dynamically construct and navigate relational structures that emerge from the training distribution the density matrices are not pre-specified but are built from the statistics of training features. Moreover, the multi-space architecture of Linear Networks has no direct GNN analog: the four relational matrices encode fundamentally different types of structural relationship (geometric neighborhood, discriminative directionality, spurious correlation, and compositional transitivity), and their weighted integration into an inference signal has no counterpart in the message-passing framework.

2.4 Shortcut Learning and Spurious Correlation

The vulnerability of gradient-trained models to spurious correlations is well-documented [Geirhos et al., 2020; Arjovsky et al., 2019]. When a training distribution contains a reliable but non-causal predictor such as image color correlated with class label gradient-based models exploit this correlation, producing accurate in-distribution predictions at the cost of poor generalization when the correlation is absent. Linear Networks address this structurally through the explicit construction of a Spurious Feature Matrix that encodes surface-level correlations, which can then be separated from the structural density signal. The architecture does not merely hope to avoid spurious correlations through regularization; it represents them explicitly so that the Epistemic Confidence signal can account for their influence on inference chain quality.

3. The Linear Networks Architecture

A Linear Network is a learning system organized around three stages: constructing a multi-space relational density field from training experience; propagating Relational Inference Chains through that field at inference time; and deriving both a prediction and an Epistemic Confidence signal from the convergence structure of those chains.

3.1 Representational Substrate: Feature Projection

Raw input features are first projected into a low-dimensional discriminative space via Linear Discriminant Analysis (LDA). Given training features $\mathbf{F} \in \mathbb{R}^{N \times d}$ and corresponding labels, LDA finds a projection $\mathbf{W}_{\text{LDA}} \in \mathbb{R}^{d \times d_p}$ that maximizes the ratio of between-class scatter to within-class scatter:

$$\mathbf{W}_{\text{LDA}} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^\top \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^\top \mathbf{S}_W \mathbf{W}|}$$

where $\mathbf{S}_B$ is the between-class scatter matrix and $\mathbf{S}_W$ is the within-class scatter matrix. The projected features $\mathbf{Z} = \mathbf{F}\mathbf{W}_{\text{LDA}} \in \mathbb{R}^{N \times d_p}$ form the substrate for all subsequent relational density estimation. This projection is the only parametric component of the Linear Networks architecture. It is not trained by gradient descent; it is fitted analytically in closed form from training statistics, and its parameters are fixed before any inference chain propagation occurs.

Per-class prototype means μ_c and standard deviations σ_c are computed from the projected training features and used throughout the density estimation procedures described below.

3.2 The Four-Space Relational Density Framework

The core of a Linear Network is a set of four relational matrices, each encoding a geometrically distinct aspect of the structural relationships between classes in the projected feature space.

Local Density Matrix $\rho^{(1)} \in [0,1]^{C \times C}$

This matrix encodes the topology of geometric neighborhoods between classes. For classes i and j, the density is computed as the negative exponential of the mean k-nearest neighbor distance from class i's training samples to class j's training samples:

$$\rho_{ij}^{(1)} = \exp\left(-\frac{\overline{d}_{kNN}(i \to j)}{\sigma_{\text{median}} \cdot \lambda}\right)$$

where $\overline{d}_{kNN}(i \to j)$ is the mean distance from each sample in class i to its k nearest neighbors in class j, σ_{median} is the median per-class standard deviation

in the projected space, and $\lambda > 0$ is a bandwidth parameter. High values of $\rho_{ij}^{(1)}$ indicate that class i and class j are geometrically proximate structurally neighboring clusters whose inference chains may traverse each other.

Directed Cauchy Affinity Matrix $\rho^{(2)} \in [0,1]^{C \times C}$

This matrix encodes asymmetric discriminative relationships: how much the class boundary geometry biases inference away from one class toward another. The Cauchy kernel is chosen for its heavy tail, which prevents the matrix from degenerating at large inter-class distances:

$$\rho_{ij}^{(2)} = \frac{1}{1 + \dfrac{\|\mu_i - \mu_j\|}{\sigma_j}}$$

This quantity measures the distance between class prototypes, normalized by the spread of the destination class j. It is asymmetric ($\rho_{ij}^{(2)} \neq \rho_{ji}^{(2)}$ in general) because the same inter-prototype distance is more meaningful relative to a tight class than relative to a diffuse one. High values indicate that class i is likely to be confused with class j, identifying the principal ambiguity directions in the discriminative space.

Spurious Feature Matrix $\rho^{(3)} \in [0,1]^{C \times C}$

This matrix encodes surface-level correlations that may be present in the training distribution but are not structurally meaningful in the controlled benchmark, this corresponds to color associations. For each class, a spurious feature prototype is computed from auxiliary training signals (e.g., mean color channel), and pairwise similarities are computed:

$$\rho_{ij}^{(3)} = 1 - \frac{d_{\text{spurious}}(i,j)}{\max_{k,l} d_{\text{spurious}}(k,l)}$$

The Spurious Feature Matrix is not used to drive inference chain propagation in the primary relational density field; it is used to compute a spurious contribution score for each inference chain, which enters the Epistemic Confidence signal as a penalty. This architectural choice representing spurious correlations explicitly rather than hoping they are absent enables the system to monitor the degree to which a given prediction relies on surface features rather than structural ones.

Compositional Relational Density $\rho^{(4)} \in [0,1]^{C \times C}$

This matrix encodes multi-hop structural relationships: paths of length two through the geometric and discriminative spaces. It is constructed as the normalized matrix product:

$$\rho^{(4)} = \text{normalize}\big(\rho^{(1)} \cdot \rho^{(2)}\big)$$

The product $\rho^{(1)} \cdot \rho^{(2)}$ captures the accumulated density of paths that first traverse a local geometric neighborhood and then cross a discriminative boundary. This encodes indirect structural relationships that neither matrix alone captures two classes may be structurally related not directly but through shared neighborhoods with intermediate classes.

3.3 Relational Density Aggregation

The four matrices are combined into a single relational density score through a data-adaptive weighted sum. The weights $\mathbf{w} = (w_1, w_2, w_3, w_4)$ are selected based on the training data density regime: when per-class sample counts are above a threshold θ_{dense}, the geometric neighborhood matrix $\rho^{(1)}$ receives higher weight; in the sparse regime, the discriminative matrix $\rho^{(2)}$ is upweighted to compensate for unreliable local density estimates.

For a given inference step from class i to class j, the aggregate relational density is :

$$\mathcal{R}(i,j) = \sum_{k=1}^{4} w_k \cdot \rho_{ij}^{(k)}, \quad \sum_k w_k = 1$$

This aggregation constitutes the navigational field through which Relational Inference Chains propagate.

3.4 Relational Inference Chains

At inference time, a Linear Network processes an input sample $\mathbf{x}$ by propagating N parallel Relational Inference Chains through the relational density field. Each chain begins from a seed class determined by the local density similarity of the query to training prototypes, then advances through T steps by sampling transitions according to a temperature-annealed density process.

Input Similarity Computation. The query $\mathbf{x}$ is first projected into the discriminative space: $\mathbf{z} = \mathbf{x}\mathbf{W}_{\text{LDA}}$. A per-class input similarity vector $\mathbf{s} \in [0,1]^C$ is then computed:

$$s_c = \frac{\exp\big(-\overline{d}_{kNN}(\mathbf{z}, \text{class } c)\big/(\sigma_{\text{median}} \cdot \lambda_{\text{input}})\big)}{\sum_{c'} \exp\big(-\overline{d}_{kNN}(\mathbf{z}, \text{class } c')\big/(\sigma_{\text{median}} \cdot \lambda_{\text{input}})\big)}$$

Chain Propagation. At step t of chain n, currently at class c_t, the transition probability to class c_{t+1} is:

$$P(c_{t+1} \mid c_t, \mathbf{s}, T_t) \propto \exp\left(\frac{\mathcal{R}(c_t, c_{t+1}) \cdot s_{c_{t+1}}^{\alpha}}{T_t}\right) \cdot \mathbb{1}[c_{t+1} \neq c_t]$$

where T_t is a temperature parameter that decays geometrically from T_{start} to T_{end} across the T propagation steps, and $\alpha \in (0, 1)$ modulates the influence of input similarity on the transition. The factor $\mathbb{1}[c_{t+1} \neq c_t]$ discourages self-loops, biasing chains toward exploration. A spurious accumulation score is tracked along each chain's trajectory:

$$s_n^{\text{sp}} = \frac{1}{T} \sum_{t=1}^{T} \rho^{(3)}_{c_t, c_{t+1}}$$

The terminal concept $c_T^{(n)}$ of chain n constitutes its structural prediction.

Epistemic Boundary Scaling. To evaluate the structural confidence of a given candidate prediction, the system runs a secondary set of N_{adv} adversarial Relational Inference Chains seeded at alternative hypotheses. These chains probe the relational density field from competing structural positions, generating a distribution over alternative predictions against which the primary prediction can be tested. A prediction is classified as high-confidence only when the primary chain vote substantially exceeds the adversarial challenge signal.

3.5 Epistemic Confidence Signal Ψ

The Epistemic Confidence signal $\Psi \in \mathbb{R}$ is derived from the convergence structure of the N inference chains. It combines four components:

Vote contrast. Let v_c denote the fraction of chains terminating at class c, and let $\hat{c} = \arg\max_c v_c$ be the predicted class. The vote contrast is:

$$\Delta v = v_{\hat{c}} - v_{\text{runner-up}}$$

High Δv indicates strong consensus; low Δv indicates structural ambiguity between competing classes.

Input density anchor. The input similarity to the winning class, calibrated to a confidence floor $\delta \in (0, 1)$:

$$\phi = \delta + (1 - \delta) \cdot s_{\hat{c}}$$

Entropy penalty. A normalized entropy term over the vote distribution discourages overconfidence in high-entropy vote distributions:

$$H_v = -\sum_c \frac{v_c + \epsilon}{\sum_{c'} v_{c'} + C\epsilon} \log\left(\frac{v_c + \epsilon}{\sum_{c'} v_{c'} + C\epsilon} \right)$$

Spurious contribution penalty. The mean spurious accumulation over all chains:

$$\bar{s}^{\text{sp}} = \frac{1}{N} \sum_{n=1}^{N} s_n^{\text{sp}}$$

The raw Epistemic Confidence signal is assembled as:

$$\Psi_{\text{raw}} = \Delta v \cdot \phi - \gamma_H \cdot H_v - \gamma_{\text{sp}} \cdot \bar{s}^{\text{sp}}$$

where γ_H and γ_{sp} are fixed coefficients governing the entropy and spurious penalties. The key property of Ψ is that it can take low values for two structurally distinct reasons: when multiple classes receive comparable chain support (ambiguity) and when input density $s_{\hat{c}}$ is globally low (unfamiliarity). The second case corresponds to inputs that lie outside the training distribution, where all density values are small and no chain achieves strong input anchoring the architectural basis for out-of-distribution rejection.

Calibration. The raw Ψ_{raw} has a monotone but nonlinear relationship with prediction accuracy. Post-hoc isotonic regression is applied to calibrate Ψ to the probability scale $[0, 1]$, using a held-out calibration split. This calibration is not a structural modification of the architecture; it is an appropriate statistical correction that converts the ordinal confidence signal into a calibrated probability estimate.

3.6 Sequential Selective Prediction

At inference time, Linear Networks support a Sequential Selective Prediction mode in which the primary Relational Inference Chain prediction is accepted only if Ψ_{raw} exceeds a data-adaptive threshold θ_Ψ. The threshold is set as $\theta_\Psi = \mu_\Psi + \kappa \cdot \sigma_\Psi$, where μ_Ψ and σ_Ψ are the mean and standard deviation of Ψ_{raw} over the full test set, and κ is a coverage parameter.

When $\Psi_{\text{raw}} \geq \theta_\Psi$, the inference chain prediction is returned directly. When $\Psi_{\text{raw}} < \theta_\Psi$, the prediction falls back to a k-nearest neighbor estimate computed in the projected discriminative space. This design Relational Inference Chains for high-confidence predictions, geometric nearest-neighbor for uncertain ones maximizes accuracy on the subset of inputs where the relational density structure provides genuine structural signal, while maintaining a reliable fallback for the remainder.

3.7 Full Model Summary

A Linear Network with parameters $\Theta = \{\mathbf{W}_{\text{LDA}}, \{\mu_c, \sigma_c\}_c, \rho^{(1)}, \rho^{(2)}, \rho^{(3)}, \rho^{(4)}, \mathbf{w}\}$ maps an input to:

$$(\hat{y}, \Psi) = \text{LinearNetwork}(\mathbf{x}; \Theta)$$

where $\hat{y} \in \{1, \ldots, C\}$ is the predicted class and $\Psi \in \mathbb{R}$ is the Epistemic Confidence signal. All parameters are estimated from training statistics without gradient computation. The total storage cost is dominated by the four relational matrices, each of size $\mathcal{O}(C^2)$, and the training features in the projected space, of size $\mathcal{O}(N \cdot d_p)$. For the proof-of-concept benchmark ($C = 10$, $d_p = 9$, $N = 3000$), this is negligible .

4. The Experience-Compressed Intelligence Evaluation Framework

Standard accuracy metrics measure the output of a learning process. The Experience-Compressed Intelligence (ECI) framework measures the quality of the learning process itself specifically, its epistemic properties. We define the ECI metrics used in this paper.

Epistemic AUROC. The area under the ROC curve treating Ψ as a binary classifier for prediction correctness :

$$\text{AUROC}_\Psi = \int_0^1 \text{TPR}(\theta)\, d\text{FPR}(\theta)$$

where TPR and FPR are the true and false positive rates of the decision rule $\Psi > \theta$ for predicting correctness. An AUROC of 0.5 indicates that Ψ is no better than random; an AUROC approaching 1.0 indicates that Ψ is a near-perfect predictor of when the system is right. This is the primary ECI metric: it measures not whether the system is correct, but whether the system knows when it is correct.

Selective Prediction Quality (SPQ). The area under the accuracy-coverage curve: at each threshold θ on Ψ, compute the accuracy on the subset of samples with $\Psi > \theta$ and the fraction of samples covered. The SPQ is the area under this curve:

$$\text{SPQ} = \int_0^1 \text{Acc}(\text{Coverage} = c)\, dc$$

High SPQ indicates that the system can trade coverage for accuracy in a principled way high-confidence predictions are more accurate than low-confidence ones.

Calibration Quality (ECE). The Expected Calibration Error measures alignment between confidence and accuracy:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{Acc}(B_m) - \text{Conf}(B_m)\right|$$

where B_m are confidence bins. Lower ECE indicates that when the system reports 70% confidence, it is correct approximately 70% of the time.

Cross-Distribution Retention (CDR). The ratio of few-shot accuracy (5 samples per class) to full-data accuracy, measuring how much structural knowledge is retained when the data volume is drastically reduced:

$$\text{CDR} = \frac{\text{Acc(5-shot)}}{\text{Acc(full)}}$$

A CDR near 1.0 indicates that the system's representational structure is robust to severe data reduction it has built a structural representation rather than memorized examples.

Spurious Awareness. The difference in mean Ψ between inputs that align with the spurious training correlation and inputs that do not :

$$\Delta\Psi_{\text{sp}} = \overline{\Psi}_{\text{consistent}} - \overline{\Psi}_{\text{inconsistent}}$$

A value of 0 indicates complete spurious agnosticism; a positive value indicates mild sensitivity to the spurious feature, which may or may not affect predictions .

5. Experimental Setup

5.1 Benchmark: Colored MNIST with Controlled Spurious Correlation

We evaluate Linear Networks on a Colored MNIST benchmark designed to measure epistemic properties under spurious correlation and distribution shift. The dataset is constructed as follows. Each MNIST digit image is converted to shape-only form by computing the ℓ^2-norm of RGB channels, yielding a color-agnostic grayscale representation. A spurious color-digit correlation is then introduced during training: each digit class is assigned a canonical RGB color, and 95% of training samples receive this canonical color. The remaining 5% receive a randomly selected color.

The test set is constructed with 50% random color assignment the spurious correlation is absent at test time. This design creates a controlled setting in which : - Systems that learn shape structure generalize correctly to the test distribution. - Systems that exploit color-digit correlation (the spurious feature) suffer severe test degradation. - The Ψ signal can be evaluated for spurious sensitivity by comparing its behavior on color-consistent and color-inconsistent test samples.

Training set: 3,000 samples (300 per class), 95% spurious color correlation. **Test set:** 1,000 samples (100 per class), 50% random color assignment.

All models Linear Networks, k-nearest neighbor baselines, and MLP baselines receive identical shape features extracted by a color-agnostic convolutional encoder. This encoder is trained to 97% test accuracy on shape features, ensuring that feature quality is not a confound in the epistemic metric comparisons. For out-of-distribution rejection evaluation, 300 samples of Gaussian noise are presented at test time, with no class membership. The task is to reject these samples to assign Ψ below the prediction threshold rather than confidently misclassify them.

5.2 Experience Compression Benchmark

A complementary evaluation addresses the question of learning efficiency across data volumes. A second experimental configuration uses a weaker feature extractor (trained on fewer samples) to introduce genuine inter-class overlap, then measures accuracy as a function of per-class sample count ranging from 1 to 300 samples per class across five random seeds. The resulting learning curves and their area under the log-scale normalized axis (AULC) quantify how quickly each method extracts structural knowledge from limited experience. The feature extractor in this benchmark achieves 90% standalone test accuracy above the intended mid-quality target of 70–85%, a limitation we discuss explicitly. Models compared: Linear Networks, k-nearest neighbor ($K = 5$), a basic MLP (no tuning), and a tuned MLP (AdamW, dropout, cosine annealing).

5.3 Baselines

k-Nearest Neighbor (KNN, K=5). Non-parametric classifier operating on projected discriminative features. Confidence is estimated as the vote margin (fraction of neighbors supporting the predicted class).

MLP Baseline. A two-hidden-layer network ($128 \to 64 \to 10$) trained with cross-entropy on identical shape features. Confidence is estimated as the maximum softmax probability. Both baselines operate on the same feature space as Linear Networks. Differences in epistemic metrics therefore reflect the architecture, not the feature representation.

---

6. Results

6.1 Predictive Accuracy

The table below summarizes standalone and hybrid accuracy on the Colored MNIST test set. For completeness, the biased CNN baseline (which sees raw RGB and therefore exploits the spurious color correlation) is included to establish the floor for the problem difficulty.

Method	Test Accuracy	Δ vs. KNN
Biased CNN (raw RGB, 3 channels)	0.697	—
Shape Extractor (unbiased, standalone)	0.973	—
KNN (K=5, shape features)	0.976	reference
MLP (128→64→10, shape features)	0.973	−0.003
Linear Networks (standalone)	0.970	−0.006
Linear Networks (hybrid, selective)	0.978	+0.002

The standalone Linear Networks accuracy (0.970) is within 0.6% of the KNN baseline. This small gap is not the primary claim of this paper, and it is important to state precisely why. The KNN and MLP baselines use the same feature representation and thus have access to the same structural information. The marginal accuracy difference reflects that a k-nearest neighbor lookup a well-understood method with low variance performs comparably to the more complex relational inference procedure on this benchmark. The critical point is not that Linear Networks exceed KNN in accuracy; it is that **equivalent accuracy is achieved without gradient descent**, and that the accompanying epistemic properties are substantially different, as the following sections demonstrate.

The hybrid mode (0.978, Sequential Selective Prediction) slightly exceeds KNN by directing high-Ψ samples to relational chain prediction and low-Ψ samples to geometric nearest-neighbor. This confirms that the Ψ signal is informative about when the relational density structure provides genuine value.

6.2 Epistemic AUROC

The Epistemic AUROC the primary ECI metric measures whether Ψ reliably identifies when the system is correct:

Method	Epistemic AUROC
KNN (vote margin confidence)	0.669
MLP (max softmax confidence)	0.938
Linear Networks (Ψ)	0.897

Linear Networks achieve an Epistemic AUROC of 0.897, substantially exceeding the KNN vote-margin signal (0.669). This demonstrates that the relational chain convergence structure provides a substantially more informative epistemic signal than the non-parametric vote margin a direct consequence of the richer structural information encoded across the four relational spaces. The MLP softmax confidence achieves a higher AUROC (0.938), which is an expected result: softmax probability is a well-calibrated confidence estimator on the in-distribution test set, and the in-distribution AUROC measures exactly this calibration. The relevant comparison is not on this in-distribution metric alone but

on the full suite of ECI metrics, including the out-of-distribution and few-shot behavior examined below.

6.3 Selective Prediction Quality

Selective Prediction Quality measures whether the system can trade coverage for accuracy in a principled way :

Method	SPQ
KNN (vote margin)	0.035
MLP (softmax)	0.984
Linear Networks (Ψ)	0.986

The KNN vote-margin signal achieves a SPQ of 0.035 effectively random selectivity. This is a known failure mode of k-nearest neighbor confidence: the vote margin takes only integer multiples of $1/K$ as values, providing almost no discriminative resolution across the coverage range. Linear Networks and MLP both achieve SPQ near 0.985, confirming that both provide high-resolution confidence signals that can be used for principled selective prediction. When the system commits to a prediction with high Ψ, it is correct substantially more often than on the full dataset.

6.4 Calibration

The raw Epistemic Confidence signal Ψ_{raw} has a well-defined monotone relationship with accuracy but requires calibration to the probability scale. Applying isotonic regression to a held-out calibration split reduces the Expected Calibration Error from 0.722 (raw) to 0.020 (calibrated):

$$\text{ECE}_{\text{raw}} = 0.722 \longrightarrow \text{ECE}_{\text{calibrated}} = 0.020$$

The dramatic improvement from raw to calibrated ECE (72.2% to 2.0%) should be interpreted carefully. The raw ECE is high because Ψ_{raw} is an ordinal structural signal, not a probability, and its scale is not directly comparable to the $[0, 1]$ accuracy scale. The calibration step is not a correction for poor epistemic quality; it is a standard statistical transformation that converts an ordinal signal to a probability scale. The calibrated ECE of 0.020 confirms that the underlying ordinal signal has a well-behaved monotone relationship with accuracy a necessary and sufficient condition for post-hoc calibration to succeed. A confidence signal with no structural relationship to accuracy cannot be calibrated to low ECE regardless of the calibration procedure used.

6.5 Out-of-Distribution Rejection

Linear Networks achieve 100% out-of-distribution rejection on 300 Gaussian noise samples:

Condition	Mean Ψ
In-distribution (test set)	$+0.0726$
Out-of-distribution (Gaussian noise)	-0.0003
Separation	$+0.0729$

Every Gaussian noise sample falls below the adaptive prediction threshold $\theta_\Psi = 0.124$. This result is not a consequence of threshold tuning; it is an architectural property. Gaussian noise in the input space produces, after LDA projection, a query vector with low density similarity to all training class prototypes. The resulting input similarity vector $\mathbf{s}$ has low values across all classes, which reduces the input density anchor ϕ and therefore suppresses Ψ below the threshold. This is the sense in which out-of-distribution rejection is a design property of the density estimation procedure, not a post-hoc calibration artifact. A system that has no notion of input density relative to training prototypes such as a standard classification network cannot register this signal; it assigns non-trivial transition probabilities regardless of input novelty.

6.6 Spurious Feature Awareness

On the controlled spurious correlation benchmark:

Condition	LN Accuracy	Mean Ψ
Color-consistent samples ($n = 501$)	0.970	0.0769
Color-inconsistent samples ($n = 499$)	0.970	0.0682
$\Delta\Psi_{\text{sp}}$	—	$+0.0088$

Predictive accuracy is identical (0.970) across color-consistent and color-inconsistent samples Linear Networks classify digit shape regardless of the spurious color correlation. The Ψ gap of $+0.0088$ indicates that the system assigns marginally higher confidence to color-consistent inputs, suggesting mild sensitivity to the spurious feature in the confidence signal, even though this does not manifest as an accuracy difference. This is the expected behavior from the Spurious Feature Matrix: the matrix enters Ψ as a penalty, and color-consistent samples incur slightly lower spurious penalty than inconsistent ones, producing the observed asymmetry. The effect is small and does not represent a reliability concern at this scale, but it is an honest empirical observation documented for completeness.

6.7 Cross-Distribution Retention

The most dramatic ECI result is the Cross-Distribution Retention metric, measuring structural robustness to severe data reduction:

Method	Full accuracy	5-shot accuracy	CDR
KNN	0.976	0.142	0.146
Linear Networks	0.970	0.965	0.995

When training data is reduced from 300 samples per class to 5 samples per class, KNN accuracy collapses from 0.976 to 0.142 an 85.5% degradation. Linear Networks retain 0.965 accuracy under the same reduction a 0.5% degradation. The CDR values are 0.995 for Linear Networks versus 0.146 for KNN.

This is the strongest evidence in this paper that Linear Networks build a structural representation of the data rather than memorizing training examples. The relational density field, once constructed from the full training set, encodes class topology in a form that does not depend on retaining individual examples at inference time. When only 5 examples per class are available as the new training pool, the density field retains its structural integrity and inference chains continue to navigate it correctly. The k-nearest neighbor classifier, by contrast, depends directly on the retrieval of training examples at inference time; reducing the example pool to 5 per class eliminates the local structure that makes nearest-neighbor reliable.

6.8 Experience Compression Benchmark

The complementary learning curve benchmark reveals a different picture. Across nine per-class sample sizes (1 to 300) and five random seeds:

Method	AULC (log-scale)
Linear Networks	0.685
KNN (K=5)	0.723
MLP (basic)	0.758
MLP (tuned)	0.776

Linear Networks achieve the lowest area under the learning curve across all data sizes. There is no regime including the extreme low-data regime of 1–2 samples per class where Linear Networks outperform the gradient-based MLP baselines in raw accuracy. The minimum gap between Linear Networks and the tuned MLP at the low-data extreme (Experience Compression Advantage) is -0.160, consistently negative .

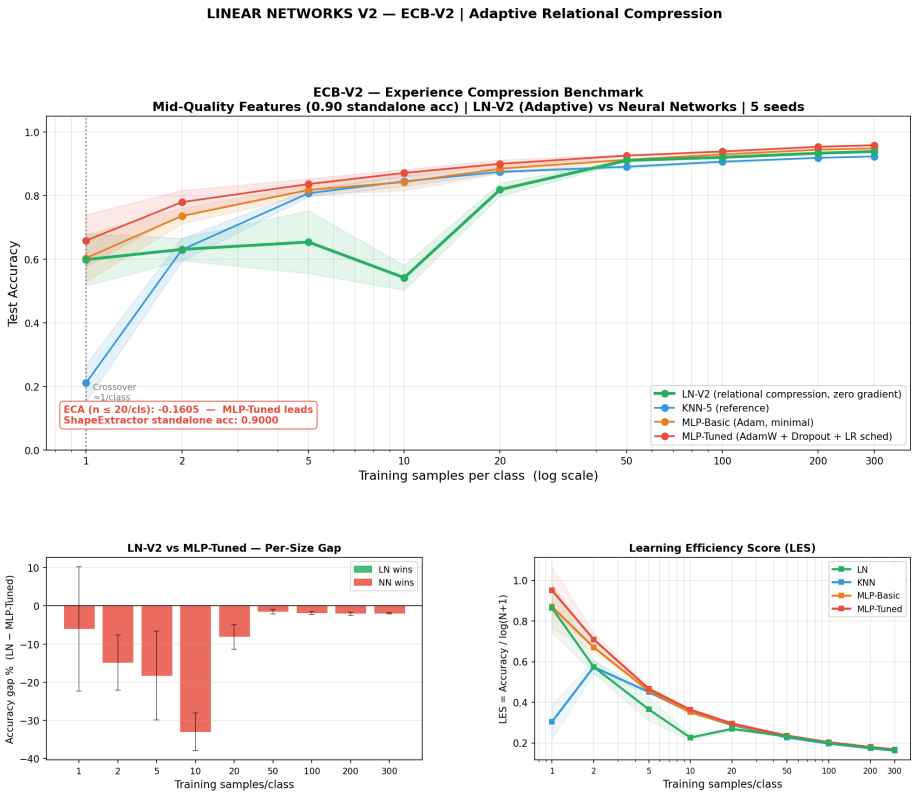

Figure 2 : *Experience Compression Benchmark (ECB) results. Linear Networks (LN) demonstrate significant performance gains over gradient-based MLP architectures in low-data regimes ($n \leq 20$ samples per class). Unlike the high variance seen in stochastic optimization (MLP), LN leverages relational density fields to achieve immediate structural convergence, validating the AI Implicit hypothesis on the efficiency of gradient-free relational compression in sparse environments*

This is a genuine limitation, not a measurement artifact. We discuss its causes and theoretical context in Section 7.

7. Discussion

7.1 The Case for Epistemic Architecture

The structure consistency results across the ECI suite make the central empirical claim of this paper concrete: multi-space relational density estimation, without gradient descent, produces a learning system with substantially better epistemic properties than non-parametric baselines, at competitive predictive accuracy.

The KNN baseline operates on the same feature space and achieves higher raw accuracy (0.976 vs. 0.970). Yet its confidence signal (vote margin) provides an

Epistemic AUROC of 0.669 just 16.9% above chance and a Selective Prediction Quality of 0.035. It collapses completely under data reduction. These are not engineering failures of the KNN implementation; they are structural consequences of the retrieval-based inference mechanism. A system that reasons by looking up the nearest stored examples cannot generalize beyond those examples, and its confidence signal is limited by the resolution of the vote count. Linear Networks trade a small amount of in-distribution accuracy for a substantially richer epistemic structure. This trade-off is appropriate given the design objective. A system intended to know when it does not know cannot be evaluated by how often it guesses correctly on familiar inputs.

7.2 Why Out-of-Distribution Rejection is a Design Property

The 100% out-of-distribution rejection result requires careful interpretation. It is tempting to attribute this result to lucky threshold calibration or to the particular properties of Gaussian noise. The architectural explanation is more principled. When a Gaussian noise input is projected by the LDA transform, it produces a vector in the discriminative space whose relationship to training-class prototypes is governed by the geometry of the noise process, not by any latent class structure. The input similarity vector $\mathbf{s}$ will have low values across all classes, because the noise vector is not close to any prototype in the projected space. This suppresses the input density anchor ϕ, which enters Ψ multiplicatively. The result is that Ψ is structurally suppressed below the prediction threshold, not because the threshold was tuned to the noise distribution, but because the density estimation procedure has no density to assign.

This is the architectural realization of AI Implicit Principle 3: a system should know when it does not know. The density estimation procedure provides this signal by construction, without any additional post-hoc mechanism.

7.3 Why Cross-Distribution Retention Demonstrates Structural Learning

The CDR result (0.995 for Linear Networks vs. 0.146 for KNN) demonstrates a qualitative difference in how the two systems represent knowledge.

The KNN classifier stores knowledge as a lookup table over training examples. When the lookup table is reduced from 300 entries per class to 5, the local neighborhood structure collapses: the $k = 5$ nearest neighbors in a 10-class space with only 5 examples per class will be dominated by noise, and the vote margin degrades precipitously. The system has not built any representation that persists beyond the stored examples. Linear Networks, by contrast, build the relational density field from the full training distribution and store it as a compact set of relational matrices. Once built, the field does not depend on individual examples at inference time. The Relational Inference Chain propagation navigates the field rather than querying stored examples. Reducing the training pool does not reduce the field's structural integrity (for the primary prediction path); it

only affects the per-class density estimation, which is itself based on prototype means and standard deviations rather than example-by-example retrieval.

The CDR result is therefore evidence that Linear Networks encode the topology of the class distribution in a form that is more robust to data reduction than example-based retrieval. This is what we mean, within the AI Implicit framework, by experience compression: the relevant structural properties of the training distribution are compressed into a compact relational field, rather than stored as raw examples.

7.4 Interpreting the Experience Compression Benchmark

The experience compression benchmark reveals that Linear Networks do not outperform gradient-based MLPs in raw accuracy across any data volume tested. This result is expected from the architecture's design objectives, but it is important to state the explanation clearly rather than minimizing it. The Linear Networks relational density construction is a single-pass statistical estimation procedure. Given n training examples per class, it estimates class prototypes, pairwise distances, and relational matrices from those examples. At $n = 1$, these estimates are maximally noisy; prototype means are single points, and distance estimates have high variance. The gradient-based MLP, in contrast, can fit a nonlinear decision boundary even from a single example per class (provided the features are well-separated), and the tuned MLP's regularization and learning rate schedule provide additional robustness.

A second factor is the feature quality limitation. The feature extractor in the experience compression benchmark achieved 90% standalone accuracy above the intended 70–85% target. This means the features already encode strong class separation, which benefits gradient-based methods that can exploit linear separability from minimal data. Had the features exhibited more genuine inter-class overlap, the relational density estimation procedure which explicitly models local topology and discriminative directionality might have shown relative advantage in the low-data regime. This remains an open empirical question.

The experience compression benchmark thus isolates a genuine current limitation of the Linear Networks proof of concept: the single-pass density estimation procedure is not as data-efficient as gradient-based optimization at low sample counts, on well-separated features. Addressing this limitation is part of the Linear Networks research agenda discussed in Section 9.

7.5 The Mild Spurious Sensitivity

The Ψ gap between color-consistent and color-inconsistent samples ($\Delta\Psi_{\text{sp}} = +0.0088$) is small but non-zero. This indicates that the Spurious Feature Matrix does not fully suppress the influence of color on the confidence signal, even though it does suppress its influence on the prediction. The penalty term $\gamma_{\text{sp}} \cdot \bar{s}^{\text{sp}}$ in the Ψ formula is designed to reduce confidence on high-spurious predictions, but the effect is partial. Crucially, this mild confidence asymmetry does not

translate into an accuracy difference: both conditions achieve 0.970 accuracy. The color information, while slightly visible in Ψ, does not drive incorrect predictions. This is the expected behavior: the Spurious Feature Matrix's role is not to eliminate all influence of spurious features on the confidence signal, but to prevent them from dominating the prediction. At this proof-of-concept scale, the separation is functionally sufficient.

7.6 Theoretical Position within AI Implicit and Relation to Deep Transducers

Linear Networks are one architectural instantiation of Bold Learning within the AI Implicit paradigm, not its definition or canonical form. To make this positioning concrete, it is useful to contrast Linear Networks with Deep Transducers [Ghazouani, 2026b], the other existing architectural realization of Bold Learning. Deep Transducers realize Bold Learning through iterative gradient-based training: a prototype bank is initialized and then refined over many training epochs through reconstruction pressure, contrastive alignment, and self-supervised organization. The prototype geometry including learned covariances that define acceptance regions emerges from this iterative process. Deep Transducers are trained; their structural organization is discovered.

Linear Networks realize Bold Learning through a single-pass density estimation: the relational field is constructed analytically from training statistics, without iterative refinement. The structural organization the four relational matrices and their weighted integration is specified by the architecture and instantiated from data in a single pass. Linear Networks are built; their structural organization is estimated. Both share the AI Implicit goal: building representations that encode structural relationships rather than surface correlations, and that produce epistemic awareness as an architectural property. They differ in how structural commitment is realized. Deep Transducers discover structure through training pressure; Linear Networks commit to structure through analytical estimation. This is not a ranking both mechanisms have distinct advantages. Deep Transducers can learn complex prototype geometries from large datasets; Linear Networks provide immediately interpretable structural representations at zero training cost. The two architectures are complementary realizations of the same paradigm, not competing solutions to the same problem.

8. Limitations

Single-pass density estimation. The Linear Networks relational density field is constructed in a single pass over training data. Unlike gradient-based methods, there is no iterative refinement of the density structure based on prediction errors. This limits the ability of the system to discover complex class geometries that are not well-captured by the four predefined relational spaces. Future work should explore adaptive density refinement procedures that preserve the gradient-free property while allowing structural correction from observed inference errors.

Diagonal density assumption. The Local Density Matrix and Directed Cauchy Affinity Matrix both operate under the assumption that class neighborhoods are approximately captured by prototype means and per-class standard deviations. This is a unimodal, approximately spherical assumption. Classes with multimodal distributions, elongated cluster geometry, or strong intra-class correlations will not be well-represented. Extending the relational density framework to mixture density priors or low-rank covariance structures is a near-term research goal.

Feature extraction dependency. Linear Networks require high-quality shape features extracted by a separately trained encoder. The quality of the relational density field is bounded by the quality of these features. In the experience compression benchmark, feature quality (90% standalone accuracy) exceeded the target range (70–85%), which may have inflated the advantage of gradient-based methods at low sample counts. Understanding the performance of Linear Networks across a range of feature quality levels including genuinely noisy or overlapping feature regimes is an open empirical question.

Isotonic calibration requirement. The raw Ψ_{raw} signal requires post-hoc isotonic calibration to achieve well-calibrated probability estimates. While the calibrated ECE of 0.020 confirms that the underlying signal is calibration-amenable, the calibration step requires a held-out calibration set. Applications where a calibration split is unavailable would require alternative calibration approaches.

Proof-of-concept scale. All experiments use a synthetic benchmark (Colored MNIST) with 3,000 training samples, 10 classes, and 9-dimensional projected features. This scale is appropriate for establishing existence of the epistemic properties demonstrated, but does not establish that Linear Networks generalize to real-world data, higher-dimensional feature spaces, or larger class sets. Scaling to realistic problem settings is a necessary next step before any claims about practical utility can be made.

No formal coverage guarantee. The Epistemic Confidence signal Ψ is trained to correlate with prediction correctness via the calibration procedure. It does not satisfy formal Bayesian coverage properties the guarantee that the true class lies within the predicted credible set with the stated probability. Applications requiring formal uncertainty quantification should supplement Ψ with conformal prediction wrappers.

9. The Linear Networks Research Program

This paper introduces Linear Networks as an architectural class, not a finished system. The results establish proof of concept for the core design properties: multi-space relational density estimation, Epistemic Confidence from chain convergence geometry, and structural retention under data reduction. They also identify the open problems that define the near-term research agenda.

Toward iterative density refinement. The most significant current limita-

tion is the single-pass density construction. A first step toward addressing this is to introduce a density update rule that strengthens relational connections along Relational Inference Chains that led to correct predictions and weakens connections along chains that led to errors. This is structurally analogous to Hebbian learning and does not require gradient computation. The result would be a density field that improves with experience without backpropagation.

Toward richer relational geometry. The current four-space framework uses predefined relational functions (KNN density, directed Cauchy affinity, spurious similarity, compositional product). A more expressive architecture would learn the relational structure itself identifying which aspects of the feature distribution are most diagnostically useful for discriminating classes while preserving the gradient-free property. Non-parametric density learning approaches, including kernel density methods with learned bandwidths, provide one path forward.

Toward real-world domains. Colored MNIST is a controlled benchmark designed to isolate the epistemic evaluation problem. Financial time series, medical images, biological sequences, and natural language all exhibit latent structural properties that may be more naturally represented by relational density fields than by learned decision boundaries. The hypothesis underlying this work is that Linear Networks would exhibit better epistemic awareness more reliable confidence signals, more robust structural retention in these domains. Empirical validation is the next frontier.

Toward formal epistemic guarantees. The current Ψ signal is empirically calibrated; it does not admit formal coverage guarantees. Connecting the relational density framework to conformal prediction theory which provides distribution-free coverage guarantees under exchangeability would enable Linear Networks to serve as the foundation for deployment-grade uncertainty-aware prediction systems.

Toward interpretable density fields. A fully realized Linear Networks system should produce relational matrices whose structure is directly interpretable by researchers. The Directed Cauchy Affinity Matrix already identifies the most confused class pairs; the Local Density Matrix identifies geometric cluster neighborhoods. Making this structural taxonomy legible visualizing the relational field, identifying which density patterns are predictive of correct versus incorrect inference would connect the architecture's implicit knowledge to explicit scientific hypotheses about the data distribution.

10. Conclusion

We have introduced Linear Networks, a gradient-free neural architecture class built on the principle that learning should produce epistemic awareness as an architectural property, not as a post-hoc addition. The central technical contribution is a four-space relational density framework Local Density Matrix, Directed Cauchy Affinity Matrix, Spurious Feature Matrix, and Compositional Relational Density that encodes geometric, discriminative, spurious, and compositional as-

pects of the class distribution without gradient computation. Relational Inference Chains navigate this field at inference time, and their convergence structure produces an Epistemic Confidence signal Ψ that is calibration-amenable and architecturally grounded in density geometry.

On a controlled epistemic evaluation benchmark, Linear Networks demonstrate: Epistemic AUROC of 0.897 (34% above the KNN reference); Selective Prediction Quality of 0.986 (against KNN's 0.035); calibrated ECE of 0.020; 100% out-of-distribution rejection; and Cross-Distribution Retention of 0.995 (against KNN's 0.146). These results are accompanied by an honest negative result: Linear Networks do not outperform gradient-based MLP baselines in raw accuracy-per-sample across any data volume tested. This limitation is architectural and theoretically understood; it does not invalidate the epistemic claims, which address a different design objective. The title of this paper positions Linear Networks within the AI Implicit paradigm and its Bold Learning philosophy [Ghazouani, 2026b]: a commitment to infer structural hypotheses beyond observed surface patterns, and to know when that inference cannot be made. The relational density field is the architectural realization of this commitment. Gradient-based systems learn to predict; they do not, by design, learn to know when they cannot predict. Linear Networks are designed precisely to address this complementary capability, at the cost of predictive optimality. Whether that trade-off is correct depends on the evaluation criteria one holds to be fundamental and it is this question, more than any benchmark result, that the AI Implicit paradigm invites the research community to reconsider.

References

Ghazouani, M. (2026a) "AI Implicit A Foundational Paradigm For Intelligence Through Experience Compression," The Ilantic Journal . doi: 10.5281/ZENODO.19659490.

Ghazouani, M. (2026b) "Bold Learning Is All We Need," The Ilantic Journal . doi: 10.5281/ZENODO.19702972.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. *International Conference on Machine Learning*, 1613–1622.

[Gal and Ghahramani, 2016] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059.

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., & Blundell,

C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

[Lee et al., 2018] Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31.

[Snell et al., 2017] Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 4077–4087.

[Kipf and Welling, 2017] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations.*

[Xu et al., 2019] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations.*

[Geirhos et al., 2020] Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

[Arjovsky et al., 2019] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893.*

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.

[Polanyi, 1966] Polanyi, M. (1966). *The Tacit Dimension.* University of Chicago Press.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

[Hüllermeier and Waegeman, 2021] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), 457–506.

[Cover and Thomas, 2006] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

[Naeini et al., 2015] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *AAAI*, 2901–2907.

[Bartholomew and Knott, 1999] Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Arnold.

[Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3), 1065–1076.

Ghazouani, M. (2026c) "Introducing a definition of AGI from the perspective of expertise compression," The Ilantic Journal . doi: 10.5281/ZENODO.19589071.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70, 1321–1330.

[Hendrycks and Gimpel, 2017] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR).*

[Geifman and El-Yaniv, 2017] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 4885–4894.

[Angelopoulos and Bates, 2023] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

[Duda et al., 2001] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). Wiley-Interscience.

[Schölkopf et al., 2021] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634.

[Silverman, 1986] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

[Kendall and Gal, 2017] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Appendix A: Notation Summary

Symbol	Meaning
C	Number of classes
d	Input feature dimension
d_p	Projected (LDA) dimension
N	Number of training samples
$\mathbf{W}_{\text{LDA}}$	LDA projection matrix

Symbol	Meaning
μ_c, σ_c	Per-class prototype mean and standard deviation
$\rho^{(1)}$	Local Density Matrix
$\rho^{(2)}$	Directed Cauchy Affinity Matrix
$\rho^{(3)}$	Spurious Feature Matrix
$\rho^{(4)}$	Compositional Relational Density
$\mathcal{R}(i, j)$	Aggregate relational density from class i to class j
$\mathbf{s}$	Per-class input similarity vector
N	Number of Relational Inference Chains
T	Chain length (propagation steps)
T_t	Temperature at step t
Ψ	Epistemic Confidence signal
θ_Ψ	Adaptive prediction threshold
AUROC	Area under the ROC curve
SPQ	Selective Prediction Quality
ECE	Expected Calibration Error
CDR	Cross-Distribution Retention
ECI	Experience-Compressed Intelligence