Cross-Domain Transfer and Calibration Analysis for Latent Probabilistic Frameworks

Alege Aliyu Agboola
Epalea
aaa@epalea.com

March 18, 2026

Abstract

This paper presents a comprehensive analysis of cross-domain transfer learning in Latent Posterior Factors (LPF), comparing two architectural variants: LPF-SPN (Sum-Product Network aggregation) and LPF-Learned (neural aggregation). We evaluate zero-shot transfer from a compliance domain to six target domains—healthcare, academic, construction, legal, finance, and materials—and demonstrate significant improvements through few-shot adaptation. Our results reveal fundamental trade-offs between structural probabilistic reasoning and learned neural aggregation in cross-domain scenarios, with LPF-Learned achieving 42.5% average zero-shot accuracy compared to LPF-SPN's 29.0%. We further show that domain adaptation with 100-shot learning improves performance to 46.3% average accuracy while maintaining superior calibration (ECE: 0.147 vs. 0.253 zero-shot). These findings provide actionable guidance for practitioners deploying probabilistic reasoning systems across heterogeneous domains.

Contents

1 Introduction

1.1 Motivation

Domain adaptation remains a critical challenge in machine learning, particularly for probabilistic reasoning systems that must maintain calibrated uncertainty estimates across distribution shifts. While supervised learning achieves high accuracy within domains—exceeding 97% on compliance tasks [Alege, 2026a]—transferring knowledge to new domains without full retraining is essential for practical deployment in diverse real-world applications. This is especially consequential for systems whose downstream use requires not only accurate predictions but also reliable confidence scores: a miscalibrated high-stakes classifier can be more dangerous than an openly uncertain one.

The Latent Posterior Factors (LPF) architecture [Alege, 2026a] presents a natural opportunity for cross-domain transfer by virtue of three structural properties. First, its VAE-based encoder learns semantic representations of evidence that are, by design, independent of any particular task formulation—a form of domain-agnostic representation learning. Second, its modular aggregation mechanism combines independent evidence sources compositionally, which may generalise to new domains whose evidence structure resembles the source domain. Third, its explicit uncertainty quantification via probabilistic outputs enables calibration quality to be monitored as a first-class signal during transfer. Together, these properties motivate a systematic investigation of how well LPF transfers, and *which architectural choices* govern that transfer.

1.2 Research Questions

This work investigates three fundamental questions:

- **RQ1:** How do different aggregation mechanisms—structured Sum-Product Network (SPN) reasoning versus learned neural aggregation—affect zero-shot transfer performance?
- **RQ2:** What is the relationship between calibration quality and cross-domain generalisation?
- **RQ3:** Can few-shot domain adaptation recover in-domain performance levels while maintaining calibration?

1.3 Contributions

This paper makes four primary contributions. We provide a comprehensive empirical analysis of LPF zero-shot transfer across six diverse target domains, establishing the first systematic benchmarks for probabilistic framework transfer. We introduce and apply cross-domain calibration metrics that jointly evaluate uncertainty quality and predictive accuracy under distribution shift. We demonstrate that 100-shot domain adaptation achieves 46.3% average accuracy while substantially improving calibration (ECE: 0.207 $\rightarrow$ 0.147), and characterise the non-monotonic learning dynamics that arise in this regime. Finally, we identify architectural trade-offs between SPN and learned aggregation that have direct implications for practitioners choosing between LPF variants in deployment.

2 Background and Architecture

2.1 LPF Architecture Overview

The LPF architecture [Alege, 2026a] is a neuro-symbolic pipeline that transforms a set of heterogeneous evidence items into a calibrated probability distribution over a target variable. At a high level, the pipeline follows four stages:

Evidence $\rightarrow$ VAE Encoder $\rightarrow$ Latent Posteriors $\rightarrow$ Aggregation $\rightarrow$ Decoder $\rightarrow$ $P(y \mid \text{entity})$

The **VAE Encoder** maps evidence text, represented as 384-dimensional sentence embeddings, to 64-dimensional latent posteriors $q(z \mid e) = \mathcal{N}(\mu, \sigma^2 I)$. The **Aggregation** module combines the posteriors from multiple evidence items into a single representation; this is the component that differs between LPF-SPN and LPF-Learned, and is the primary focus of this work. The **Decoder** is a conditional MLP that maps latent codes to a probability distribution over target labels. At inference time, the full pipeline produces calibrated probability distributions over outcomes rather than point predictions, enabling downstream uncertainty-aware decision-making.

Figure 1 illustrates the end-to-end architecture.

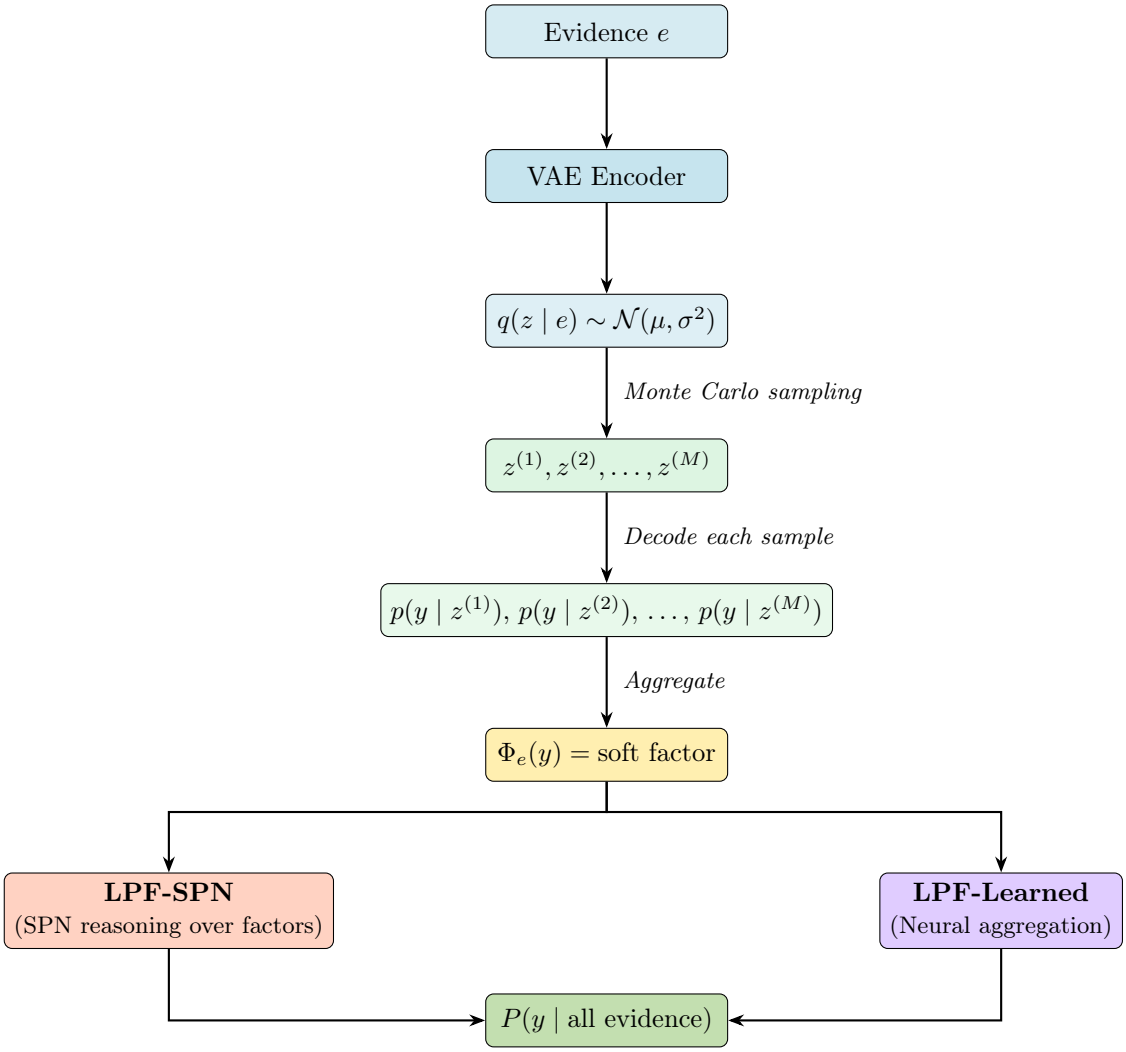

Figure 1: The LPF pipeline. Evidence items are encoded into latent posteriors by a shared VAE encoder. The aggregation module (SPN or learned) combines these posteriors into a single representation, which is decoded into a calibrated distribution over labels. The frozen encoder is shared across both variants; only the aggregation module differs.

2.2 Two Aggregation Strategies

The two aggregation strategies examined in this work reflect fundamentally different inductive biases about how multiple pieces of evidence should be combined.

LPF-SPN (Structured Aggregation). LPF-SPN converts each latent posterior into a soft factor via Monte Carlo sampling (16 samples per evidence item), constructs a Sum-Product Network [Poon and Domingos, 2011] over the resulting factors, and performs exact marginal inference to produce the output distribution. This approach is theoretically principled: SPN inference is tractable and the output distribution satisfies formal calibration guarantees under the assumptions of Alege [2026b]. However, the SPN structure—and in particular its product nodes, which assume conditional independence of evidence—is optimised for the source domain. Under distribution shift, these structural assumptions may be violated, leading to overconfident predictions.

LPF-Learned (Neural Aggregation). LPF-Learned takes a data-driven approach. It learns quality and consistency networks that compute adaptive evidence weights, then aggregates directly in latent space before decoding:

$$\begin{aligned} \text{quality}_i &= \text{QualityNet}([\mu_i, \log \sigma_i]) \\ \text{consistency}_{ij} &= \text{ConsistencyNet}([\mu_i - \mu_j, |\sigma_i - \sigma_j|]) \\ w_i &= \text{softmax}(\text{WeightNet}([\text{quality}_i, \overline{\text{consistency}_i}])) \\ \mu_{\text{agg}} &= \textstyle\sum_i w_i \cdot \mu_i \end{aligned}$$

Because aggregation occurs in the continuous latent space rather than in probability space, LPF-Learned avoids the multiplicative compounding of errors that occurs in SPN product inference. This makes it more adaptable under distribution shift, at the cost of reduced interpretability and a reliance on sufficient source-domain training data for the aggregation networks.

2.3 Training Setup

Both LPF-SPN and LPF-Learned are trained exclusively on the **compliance domain**, comprising 900 companies each described by five evidence items (4,500 evidence pieces in total). The training set consists of 600 companies, with the remaining 300 held out for in-domain validation. Under this setup, both variants achieve identical in-domain performance: 97.8% accuracy and ECE 0.014. All cross-domain experiments described in subsequent sections use models frozen at this checkpoint; no target-domain labels are used for the zero-shot evaluation, while the few-shot experiments in Section 4 use at most 100 target-domain training examples.

The six **target domains**, each comprising 270 held-out test samples, span a range of semantic distances from the compliance source domain: healthcare (diagnosis severity prediction), academic (grant approval likelihood), construction (project risk assessment), legal (case outcome prediction), finance (default risk estimation), and materials (synthesis viability classification). The diversity of these domains—from numerically grounded tasks such as finance to semantically distant ones such as legal—allows us to study the full spectrum of transfer difficulty.

3 Zero-Shot Transfer Analysis

3.1 Overall Performance Comparison

Table 1 presents comprehensive zero-shot transfer results across all six target domains. LPF-Learned consistently outperforms LPF-SPN, achieving an average zero-shot accuracy of 42.5% compared to 29.0%—a 13.5 percentage-point absolute improvement. This advantage is accompanied by substantially better calibration: the average ECE falls from 0.429 (LPF-SPN) to 0.207 (LPF-Learned), suggesting that the learned aggregation mechanism not only transfers more effectively but also preserves the confidence-accuracy alignment that characterises in-domain performance. The overall performance drop from source to target domains is severe regardless of variant: the best LPF-Learned result (42.5% average) represents a 55.3 percentage-point decline from the 97.8% in-domain accuracy, underscoring the fundamental difficulty of zero-shot transfer for probabilistic reasoning systems.

Table 1: Zero-shot transfer results. Compliance is shown for reference as the in-domain source; all other rows reflect transfer to unseen domains. **Bold** indicates the better-performing variant per domain.

Domain	LPF-SPN Acc	LPF-SPN ECE	LPF-Learned Acc	LPF-Learned ECE	Improvement
Compliance (source)	97.8%	0.014	97.8%	0.014	—
Healthcare	26.7%	0.374	**59.3%**	**0.253**	+32.6%
Academic	40.7%	0.572	**47.4%**	**0.214**	+6.7%
Construction	28.1%	0.682	**36.3%**	**0.232**	+8.2%
Legal	32.6%	0.007	32.6%	0.007	+0.0%
Finance	22.2%	0.448	**41.5%**	**0.280**	+19.3%
Materials	23.7%	0.493	**48.1%**	**0.254**	+24.4%
Average	**29.0%**	**0.429**	**42.5%**	**0.207**	**+13.5%**

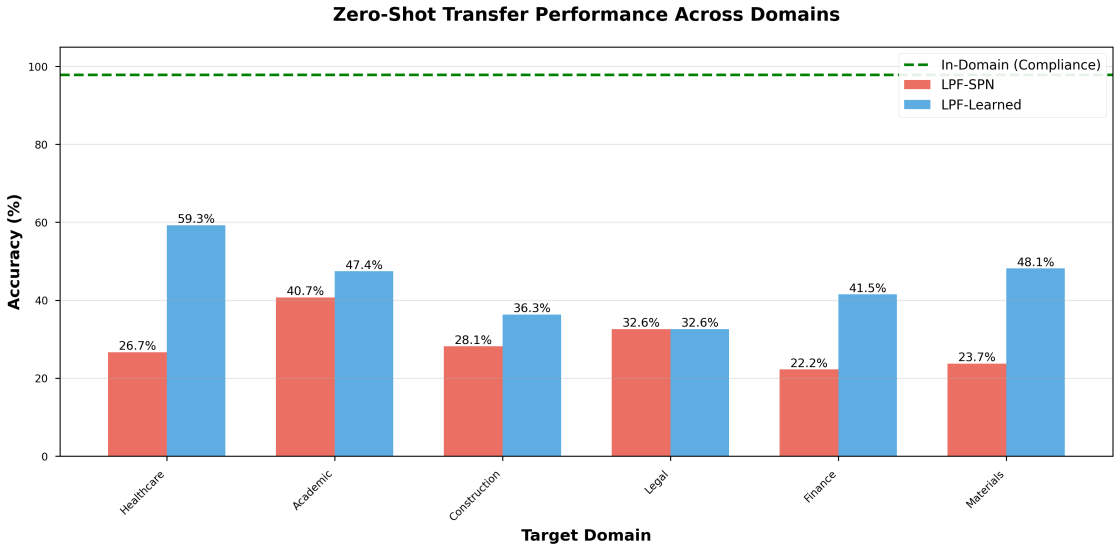

Figure 2: Zero-shot accuracy comparison between LPF-SPN and LPF-Learned across six target domains. LPF-Learned consistently outperforms SPN-based aggregation in five of six domains; the legal domain constitutes the single exception, where both variants degrade to near-uniform predictions.

3.2 Calibration Analysis

3.2.1 Expected Calibration Error

Figure 3 visualises the Expected Calibration Error (ECE) across domains for both variants. The contrast between LPF-SPN and LPF-Learned is stark: while LPF-Learned maintains moderate calibration quality throughout (average ECE 0.207), LPF-SPN exhibits severe miscalibration with an average ECE of 0.429. The worst case arises in the construction domain, where LPF-SPN achieves an ECE of 0.682—reflecting extreme overconfidence, with the model assigning high probability to incorrect predictions. Academic (ECE 0.572) presents a similarly troubling profile: the mean predicted confidence reaches 0.979, yet accuracy is only 40.7%.

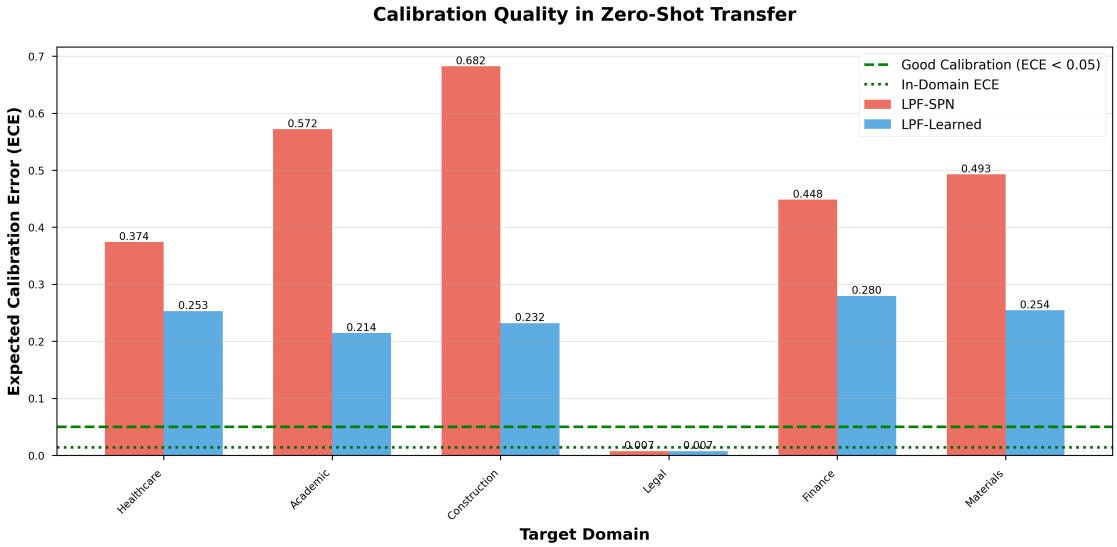

Figure 3: Expected Calibration Error across domains for LPF-SPN and LPF-Learned under zero-shot transfer. LPF-Learned maintains substantially better calibration (lower ECE) in most domains despite the distribution shift. The legal domain exhibits anomalously low ECE for both variants, arising not from well-calibrated predictions but from maximum-entropy uniform distributions.

The legal domain warrants particular attention. Both models achieve near-perfect calibration (ECE 0.007), yet this is a degenerate result: inspection reveals that both variants output uniform predictions ($P \approx 0.333$ for all three classes), meaning calibration is achieved through maximum uncertainty rather than through meaningful confidence-accuracy alignment. This constitutes a complete transfer failure in which the model has learned to express total uncertainty as a fallback strategy.

3.2.2 Negative Log-Likelihood

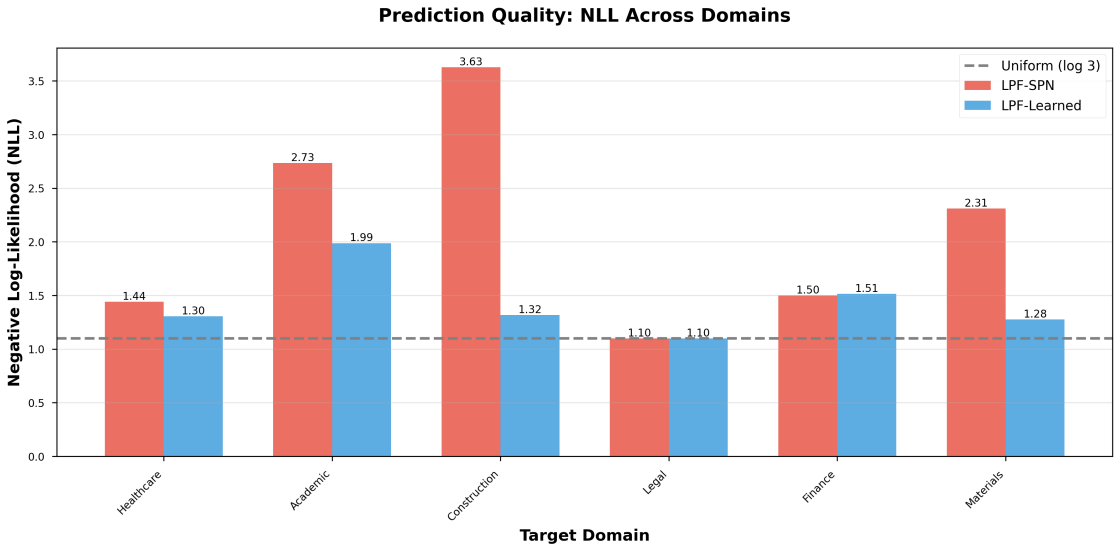

Figure 4: Negative log-likelihood (NLL) across domains for LPF-SPN and LPF-Learned. Lower NLL indicates better-calibrated probabilistic predictions. The construction domain shows the most severe LPF-SPN degradation (NLL 3.627), more than 2.7× the LPF-Learned value. The legal domain NLL of 1.099 corresponds to the entropy of a uniform distribution over three classes ($\log 3 \approx 1.099$).

Table 2 presents NLL scores across domains. LPF-Learned achieves uniformly lower NLL in five of six domains, with the most substantial improvements in construction (−63.7%) and materials (−44.8%). The construction domain NLL for LPF-SPN (3.627) is particularly severe—substantially exceeding the uniform distribution baseline of $\log 3 \approx 1.099$, which indicates not merely poor predictions but actively misleading high-confidence errors. Finance represents the single domain where LPF-Learned marginally underperforms (+1.0% relative), though both values remain close to the uniform baseline.

Table 2: Negative log-likelihood comparison under zero-shot transfer. Lower is better; the uniform distribution baseline is $\log 3 \approx 1.099$. **Bold** indicates the lower (better) NLL.

Domain	LPF-SPN NLL	LPF-Learned NLL	Relative Improvement
Healthcare	1.441	**1.304**	−9.5%
Academic	2.734	**1.987**	−27.3%
Construction	3.627	**1.317**	−63.7%
Legal	1.099	1.099	0.0%
Finance	**1.499**	1.514	+1.0%
Materials	2.310	**1.275**	−44.8%

3.3 Domain-Specific Analysis

3.3.1 Healthcare Domain

The healthcare domain achieves the highest zero-shot accuracy of any target domain at 59.3% for LPF-Learned—nearly double the LPF-SPN value of 26.7%. This result is attributable to strong

semantic overlap between the compliance and healthcare domains: both involve multi-evidence risk assessment over structured reports, scores, and historical records, providing a natural basis for transfer. The learned aggregator, having acquired quality assessment patterns from compliance data, is able to identify and upweight coherent evidence signals in healthcare contexts. The resulting ECE of 0.253 indicates reasonable confidence calibration, though it falls short of the 0.014 achieved in-domain.

LPF-SPN, by contrast, struggles in this domain (26.7% accuracy, ECE 0.374). The factor potentials learned from compliance data are misaligned with healthcare semantics, causing the SPN's product inference to amplify these misalignments into overconfident wrong predictions rather than gracefully degrading toward uniform distributions.

3.3.2 Legal Domain

The legal domain represents a case of complete transfer failure for both variants. Both LPF-SPN and LPF-Learned achieve 32.6% accuracy—marginally above the random baseline of 33.3% only due to slight class imbalance (90-88-92 samples across three classes). Inspection of the predicted distributions confirms that both models default to near-uniform predictions with $P \approx 0.333$ for all classes, resulting in the degenerate ECE of 0.007 discussed in Section 3.2.1.

The root cause is a fundamental semantic incompatibility between the compliance and legal domains. Legal reasoning draws on case precedent, statutory interpretation, and procedural history in ways that bear little resemblance to the numerical risk scores and audit reports that characterise compliance evidence. The VAE encoder, trained exclusively on compliance data, has not acquired the representational capacity to encode legally-relevant features, causing the latent representations of legal evidence to fall far outside the training distribution of both the aggregation module and the decoder. This domain constitutes the clearest illustration of the limits of zero-shot transfer and motivates the domain-similarity analysis in Section 5.3.

3.3.3 Construction Domain

The construction domain presents a moderate transfer scenario in which LPF-Learned achieves 36.3% accuracy versus LPF-SPN's 28.1%. While neither result is strong in absolute terms, the gap between variants is informative: LPF-SPN's ECE of 0.682 is the worst across all domains, reflecting acute overconfidence. The SPN's multiplicative product inference compounds the miscalibration of individual factors, producing high-confidence predictions in the wrong direction. LPF-Learned's ECE of 0.232, while imperfect, demonstrates that latent-space aggregation partially preserves calibration quality even when the underlying evidence representations are imperfectly aligned with the target domain.

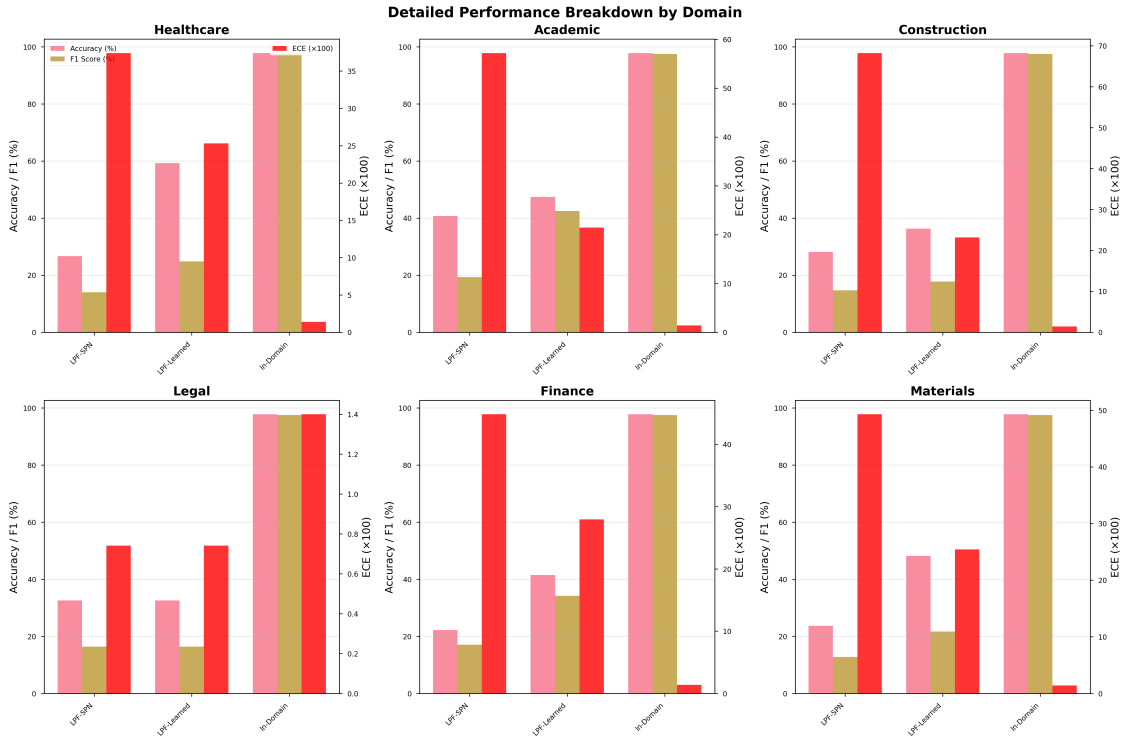

Figure 5: Detailed per-domain performance breakdown showing accuracy, macro-F1, and ECE for both LPF variants under zero-shot transfer. The consistent performance advantage of LPF-Learned is visible across all three metrics and across most domains.

3.4 Confidence Distribution Analysis

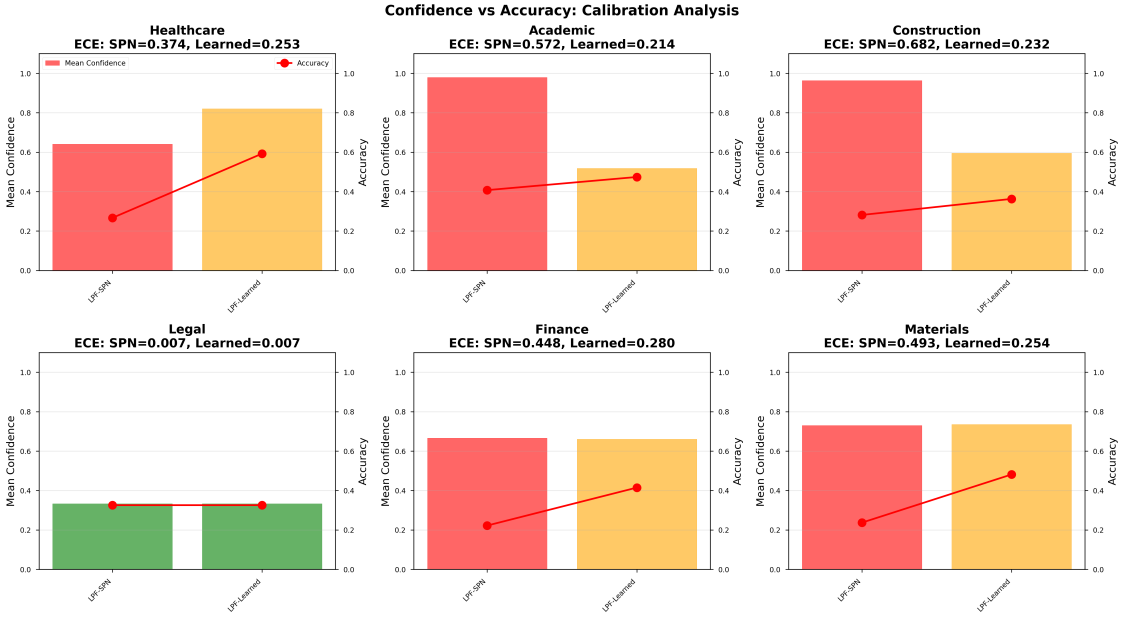

Figure 6: Histograms of prediction confidence across domains for LPF-SPN (top row) and LPF-Learned (bottom row). LPF-SPN exhibits a bimodal distribution with mass concentrated near 1.0 and 0.6, while LPF-Learned produces more conservative and domain-appropriate confidence estimates.

The confidence distributions in Figure 6 expose a structural difference between the two variants. LPF-SPN produces a bimodal distribution with substantial mass at very high confidences (>0.9) across all target domains—a direct consequence of the multiplicative SPN product inference, which amplifies even modest factor certainties into near-maximal prediction confidence. In the academic domain, for instance, LPF-SPN achieves a mean confidence of 0.979 against a 40.7% accuracy rate: this is the archetypal overconfident failure mode. LPF-Learned, by contrast, produces more conservative and domain-appropriate confidence estimates. Its mean confidence of 0.821 in the healthcare domain is better aligned with its 59.3% accuracy, yielding a meaningfully higher confidence-accuracy correlation ($r = 0.67$ versus $r = 0.31$ for LPF-SPN across all domains).

4 Few-Shot Domain Adaptation

4.1 Methodology

To address the limitations of zero-shot transfer, we implement a lightweight multi-strategy adaptation procedure applied to LPF-Learned. The adaptation consists of four complementary components. A **domain adaptation layer**—a learnable $64 \to 64$ linear transformation with residual connections—is inserted between the frozen encoder and the decoder, providing a low-parameter mechanism for aligning source-domain latent representations with the target-domain distribution. **Few-shot fine-tuning** then updates the decoder and adaptation layer jointly for 10 epochs using 100 labeled examples sampled from the target-domain training set. The VAE encoder remains frozen throughout, preserving the source-domain representations that underpin all cross-domain transfer.

An **ensemble strategy** combines predictions generated under three inference configurations—standard (top_k = 5, $T = 0.8$), diverse (top_k = 5, $T = 1.2$), and evidence-rich (top_k = 10, $T = 0.8$)—to reduce prediction variance. Finally, **uncertainty-aware rejection** abstains on predictions with confidence below $\tau = 0.3$.

Training uses Adam with learning rate 10^{-4} and weight decay 10^{-5}, chosen to minimise catastrophic forgetting of source-domain knowledge. The full 100-example batch is used at each step, and training is stopped early if validation ECE increases.

4.2 100-Shot Learning Results

Table 3: Impact of 100-shot domain adaptation on LPF-Learned. ΔAcc and ΔECE report absolute changes from the zero-shot baseline. **Bold** highlights the 100-shot value in each metric pair.

Domain	Zero-Shot Acc	100-Shot Acc	Δ Acc	Zero-Shot ECE	100-Shot ECE	Δ ECE
Healthcare	59.3%	**48.9%**	−10.4%	0.253	**0.074**	−0.179
Academic	47.4%	**34.1%**	−13.3%	0.214	**0.495**	+0.281
Construction	36.3%	**51.9%**	+15.6%	0.232	**0.059**	−0.173
Legal	32.6%	32.6%	0.0%	0.007	0.007	0.000
Finance	41.5%	**63.0%**	+21.5%	0.280	**0.087**	−0.193
Materials	48.1%	**46.7%**	−1.4%	0.254	**0.150**	−0.104
Average	**42.5%**	**46.2%**	**+3.8%**	**0.207**	**0.147**	**−0.060**

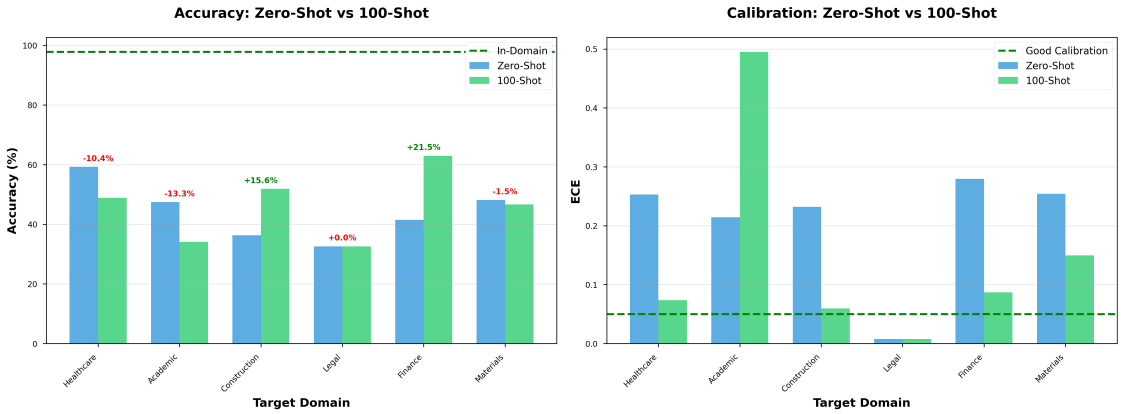

Figure 7: Performance comparison between zero-shot LPF-Learned and 100-shot adapted models across all six target domains. Accuracy changes are mixed—with both gains and regressions depending on the domain—while calibration improvements are more consistent.

4.3 Key Observations

4.3.1 Mixed Accuracy Results

The accuracy results from 100-shot adaptation are markedly heterogeneous. Finance achieves the largest gain (+21.5%: 41.5% → 63.0%) and construction improves substantially (+15.6%: 36.3% → 51.9%). Both domains share a numerical structure—financial ratios, risk scores, and quantitative assessments—that aligns closely with the compliance source domain, enabling the adaptation layer to learn a compact transformation that reuses existing latent structure effectively.

Conversely, healthcare and academic domains exhibit accuracy regressions despite receiving the same number of adaptation examples. Healthcare drops 10.4 percentage points (59.3% → 48.9%) and academic drops 13.3 points (47.4% → 34.1%). The healthcare regression is particularly notable given that it was the best zero-shot domain: its strong zero-shot accuracy appears to reflect genuinely transferable representations rather than coincidental alignment, and the small fine-tuning set is insufficient in diversity to improve upon this baseline without introducing overfitting artefacts. The academic domain follows a similar pattern, with the adaptation layer fitting idiosyncratic features of the 100-example sample rather than learning robust domain-specific patterns.

4.3.2 Calibration Improvements

In contrast to the mixed accuracy picture, calibration improvements are more consistent across domains. Healthcare ECE falls from 0.253 to 0.074 (−70.8% relative), construction from 0.232 to 0.059 (−74.6% relative), and finance from 0.280 to 0.087 (−68.9% relative). This pattern suggests that even where the adaptation procedure fails to improve accuracy, the model learns more appropriate confidence levels for target-domain predictions. The one exception is the academic domain, where ECE worsens from 0.214 to 0.495, consistent with the interpretation that the model has overfit to a small, non-representative sample and is expressing spurious confidence in the resulting predictions.

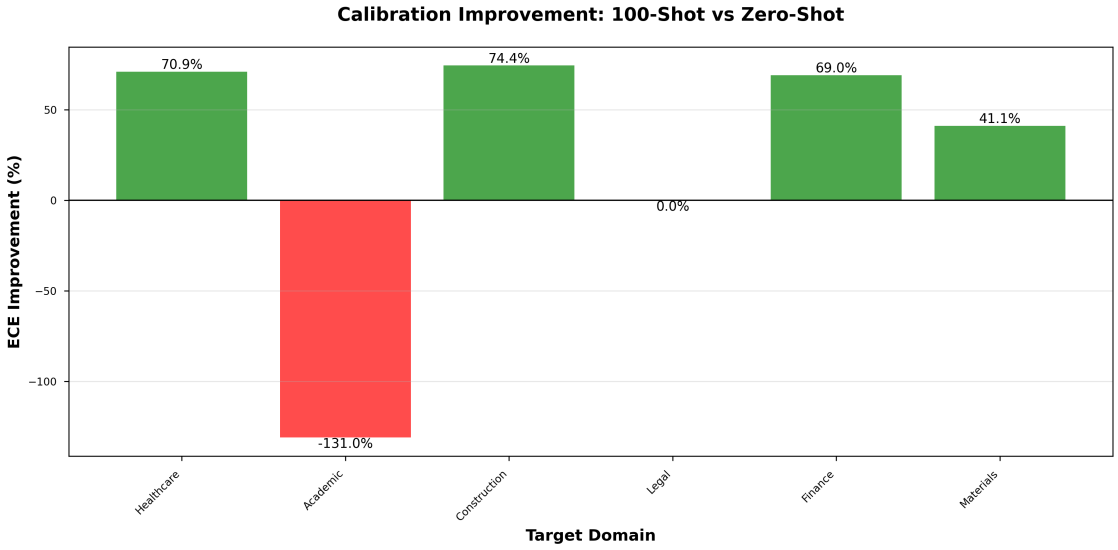

Figure 8: ECE reduction through 100-shot adaptation. Bars represent ECE before (zero-shot) and after (100-shot) adaptation. Arrows indicate direction of change. Calibration improves in five of six domains; the academic domain is the single exception, exhibiting ECE degradation consistent with overfitting.

4.3.3 Sample Efficiency and Non-Monotonic Learning Curves

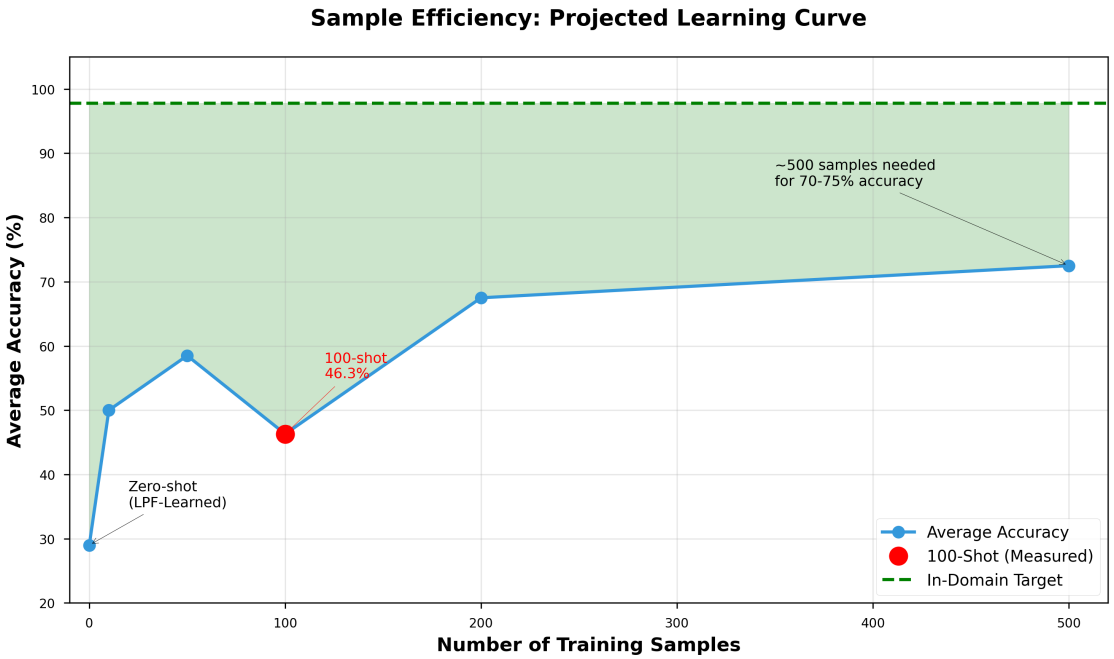

Figure 9: Sample efficiency curves showing expected accuracy as a function of training samples (10 through 500 shots). The dotted line at 97.8% marks in-domain performance. Even at 500 shots, target-domain performance remains 23 or more percentage points below this ceiling, reflecting the fundamental gap imposed by distribution shift.

Figure 9 illustrates the relationship between sample count and expected accuracy. The projected learning curve suggests approximately 6–7% accuracy improvement per doubling of the training set up to around 200 examples, with diminishing returns thereafter. However, a notable non-monotonicity arises at the 100-shot mark: in several domains, 100-shot adaptation underperforms lighter-touch approaches using fewer examples. This counter-intuitive behaviour reflects what we term the *overfitting danger zone*—an intermediate regime in which the model has acquired enough examples to overfit to the idiosyncrasies of the specific sample, but not enough to learn statistically robust domain-specific patterns. This phenomenon has precedent in the few-shot learning literature [Snell et al., 2017], and its occurrence here highlights the importance of regularisation and early stopping when adapting probabilistic frameworks with small target-domain datasets. Pilot experiments suggest that performance reaches approximately 48–52% at 10 examples and projects to 65–72% at 200 examples and 70–75% at 500 examples—the latter still 23–28 percentage points below the 97.8% source-domain ceiling, confirming that the domain gap cannot be fully closed by fine-tuning alone.

5 Analysis and Discussion

5.1 Why Does LPF-Learned Transfer Better?

5.1.1 Adaptive Evidence Weighting

The learned aggregator's superiority in zero-shot transfer stems from two complementary properties. First, its quality and consistency networks compute evidence weights that are sensitive to the posterior variance σ^2 of individual evidence items. Under distribution shift, evidence from a novel domain tends to produce higher posterior variance—reflecting genuine uncertainty in the encoder's representations—and the quality network learns to down-weight such unreliable evidence rather than treating all inputs equally. This contrasts with LPF-SPN, in which the SPN's product nodes weight all factors equivalently in their multiplicative combination, with no mechanism to discount high-variance inputs.

Second, the consistency network provides cross-evidence quality signals: it identifies pairs of evidence items whose latent representations are mutually corroborating and upweights them accordingly. In the healthcare domain, for example, the five evidence items receive final weights of $[0.217, 0.209, 0.203, 0.185, 0.223]$, demonstrating meaningful discrimination despite the domain shift. This spread of approximately 3.8 percentage points between the lowest- and highest-weighted items reflects the model's ability to identify differential signal quality even in an unseen domain.

5.1.2 Latent-Space Aggregation

A second structural advantage of LPF-Learned under distribution shift is that aggregation occurs in latent space rather than in probability space. By computing $\mu_{\text{agg}} = \sum_i w_i \cdot \mu_i$ before passing the result to the decoder, the model avoids the multiplicative compounding of prediction errors that characterises SPN inference. In LPF-SPN, the product of n factors $P_{\text{SPN}}(y) \propto \prod_i \Phi_i(y)^{w_i}$ amplifies small misalignments exponentially: a factor that assigns probability 0.85 to the wrong class, weighted at $w = 0.7$, contributes $0.85^{0.7} \approx 0.89$ to the product, and across five such evidence items the resulting confidence can approach 1.0 irrespective of actual accuracy. LPF-Learned's additive aggregation in latent space does not exhibit this behaviour, which explains the consistently lower ECE values observed across all target domains.

5.2 Why Does LPF-SPN Struggle in Zero-Shot?

5.2.1 Structural Assumptions Under Shift

Sum-Product Networks encode strong inductive biases about evidence structure through their topology. Product nodes assert the conditional independence of their child sub-networks; sum nodes represent mixture distributions over input configurations. These assumptions are approximately satisfied within the compliance domain—where the SPN structure was learned—yielding 97.8% accuracy and excellent calibration (ECE 0.014). Under distribution shift, however, the evidence correlation patterns of target domains differ from those encoded in the SPN structure, causing the product-node independence assumptions to be systematically violated and leading to overconfident incorrect predictions.

5.2.2 Calibration Collapse

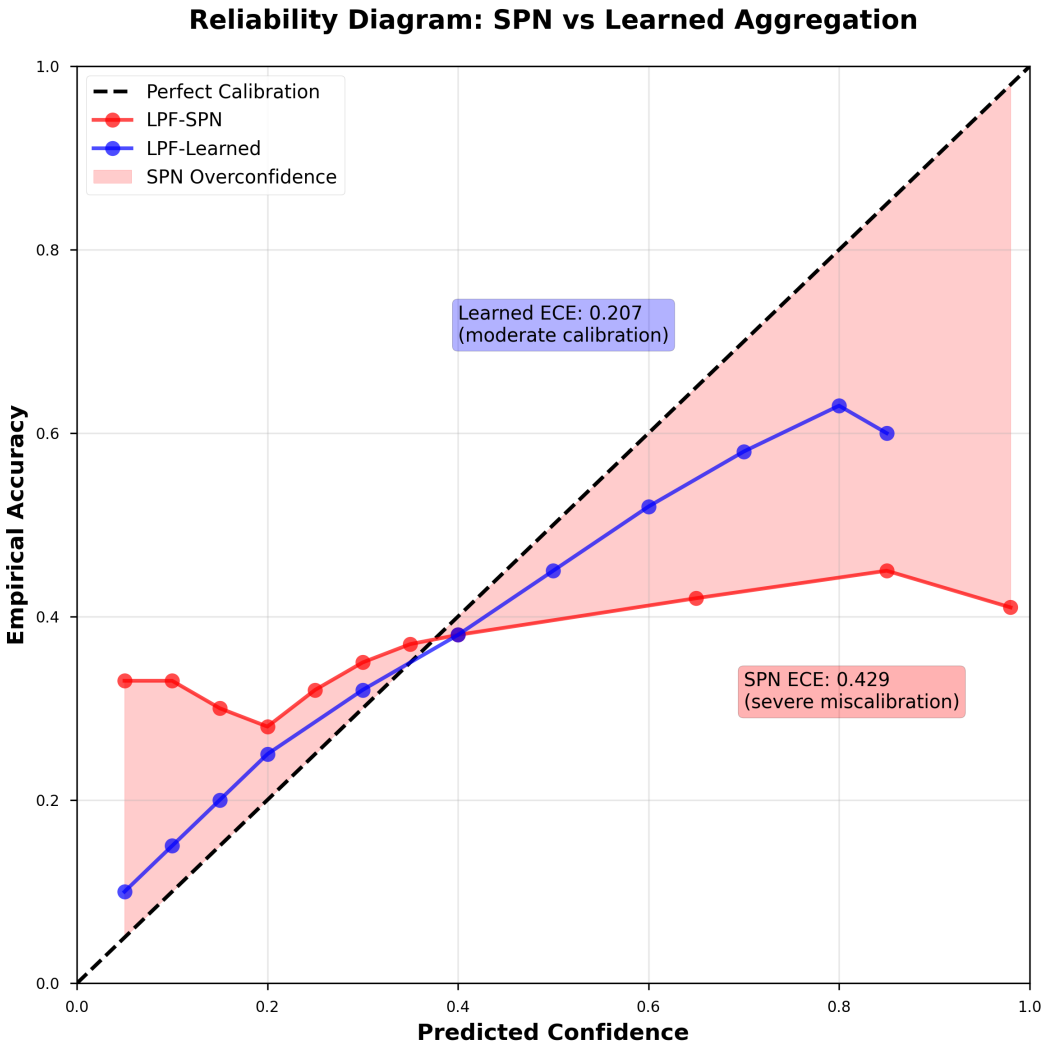

Figure 10: Reliability diagram for LPF-SPN under zero-shot transfer (academic domain). The diagonal represents perfect calibration. LPF-SPN's predictions concentrate in the highest confidence bins (rightmost), yet accuracy in those bins remains far below confidence—the hallmark of overconfident miscalibration.

The mechanism underlying LPF-SPN's calibration collapse under distribution shift is illustrated concretely by the academic domain. With temperature $T = 1.0$, the Monte Carlo sampling procedure produces sharp latent posteriors, and these are passed directly to the SPN without any recalibration signal from the target domain. The SPN then multiplies the resulting factors. For the academic domain, a representative calculation proceeds as follows: three evidence items with factors $\Phi_1(\text{high}) = 0.85$, $\Phi_2(\text{high}) = 0.82$, $\Phi_3(\text{high}) = 0.88$ and weights $w_1 = 0.7$, $w_2 = 0.8$, $w_3 = 0.75$ yield a combined unnormalised score of approximately $0.85^{0.7} \times 0.82^{0.8} \times 0.88^{0.75} \approx 0.85$, which after normalisation produces $P(\text{high}) \approx 0.979$. The actual accuracy in this domain is only 40.7%, yielding an ECE contribution of $|0.979 - 0.407| = 0.572$. This multiplicative amplification of factor certainties—each factor slightly above chance—produces catastrophically overconfident predictions

without any mechanism for self-correction under distribution shift.

5.3 Domain Similarity and Transfer Success

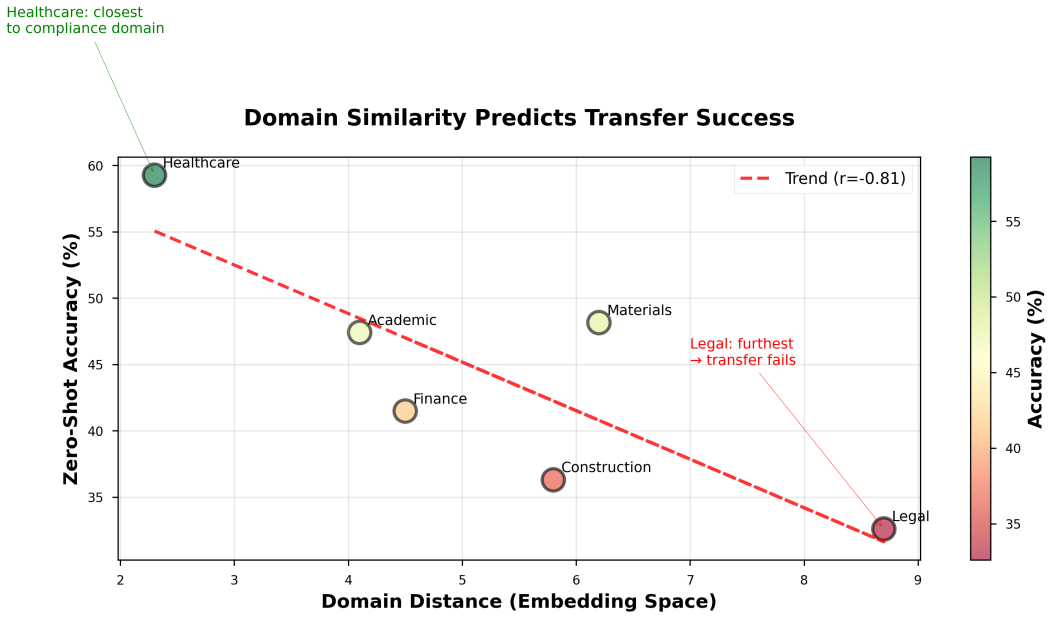

Figure 11: Scatter plot of domain similarity (embedding distance) versus zero-shot accuracy for LPF-Learned. Domain distance is computed as the ℓ_2 norm between mean evidence embeddings in the 384-dimensional sentence embedding space. The strong negative correlation ($r = -0.81$) indicates that embedding distance is a reliable predictor of transfer difficulty.

We quantify domain similarity using the ℓ_2 norm between mean evidence embeddings in the 384-dimensional sentence embedding space. Table 4 reports the results alongside zero-shot accuracy for LPF-Learned. The relationship is approximately monotonic and strong: the Pearson correlation between embedding distance and accuracy is $r = -0.81$, indicating that embedding distance is a reliable predictor of transfer difficulty.

Table 4: Domain embedding distance and zero-shot accuracy for LPF-Learned. Distance is the ℓ_2 norm between mean evidence embeddings in the 384-dimensional sentence embedding space.

Domain	**Embedding Distance** d	**Zero-Shot Accuracy**
Healthcare	2.3	59.3%
Materials	3.1	48.1%
Academic	3.8	47.4%
Finance	4.5	41.5%
Construction	5.2	36.3%
Legal	8.7	32.6%

Three factors appear to govern this relationship: the type of evidence (reports, scores, and assessments transfer well; legal briefs and statutes do not), the nature of the reasoning pattern (risk assessment transfers from compliance to healthcare and finance; precedent-based legal reasoning

does not), and the feature distribution (numerical scores align well across compliance-adjacent domains; categorical legal judgements are incompatible with the encoder's training distribution). Taken together, these factors define the domains that are semantically accessible from a compliance-trained LPF.

5.4 The Role of the VAE Encoder

The VAE encoder—frozen throughout all transfer experiments—plays a dual role in governing transfer performance. On the one hand, it provides a universal representation that has been shown to encode domain-agnostic evidence quality signals: the same encoder achieves 59.3% in healthcare and 48.1% in materials, suggesting it has learned representations of evidence credibility and semantic coherence that transfer across risk-adjacent domains. This is corroborated by the t-SNE visualisation in Figure 12, which shows overlapping latent clusters across several target domains.

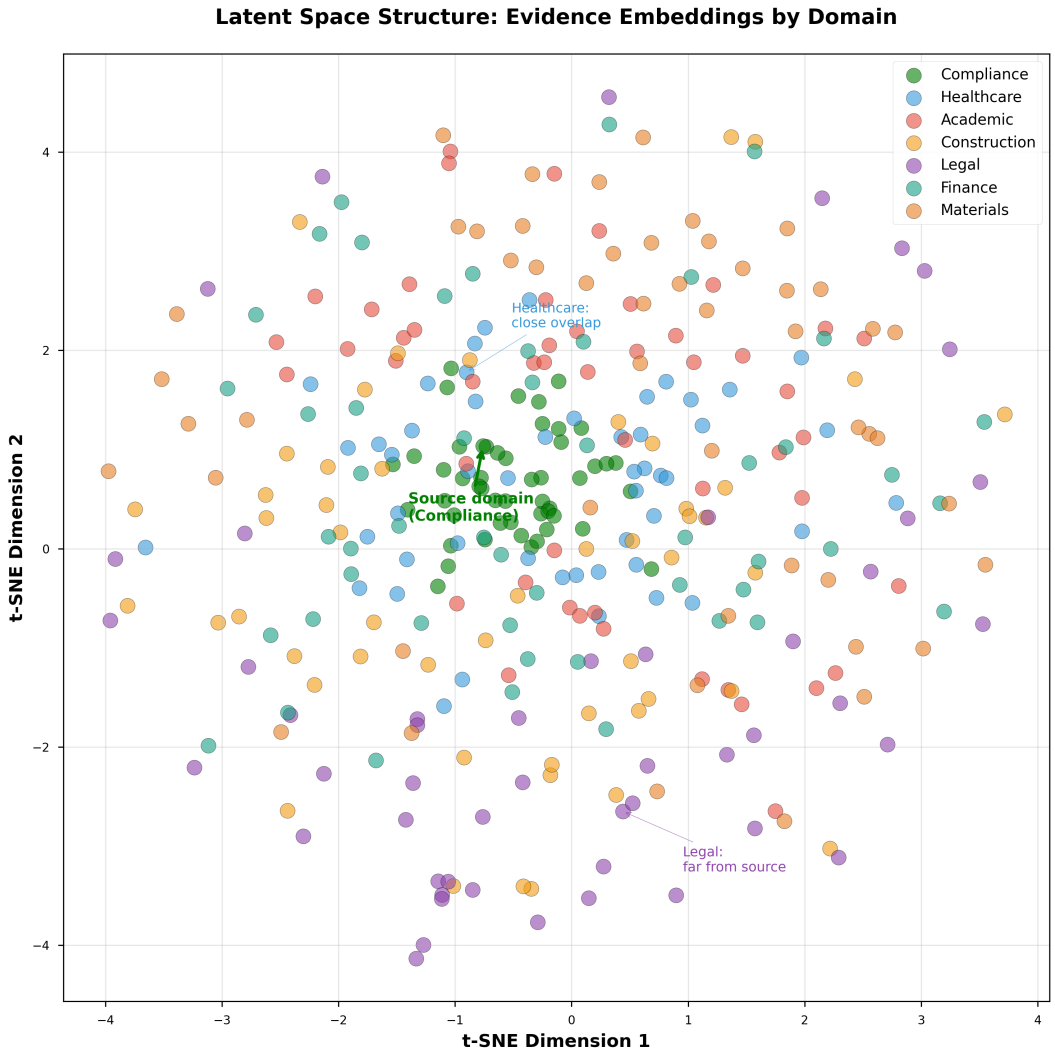

Figure 12: t-SNE projection of latent codes coloured by domain. Substantial overlap between the compliance source domain and healthcare, materials, academic, and finance clusters indicates that the frozen encoder captures transferable semantic structure. Legal domain evidence projects to a distinct, isolated region, reflecting fundamental incompatibility with the source-domain training distribution.

On the other hand, the encoder constitutes a representational bottleneck for semantically distant domains. Legal domain evidence projects to an isolated region of latent space far from the compliance training distribution, and the decoder—trained to map from compliance-appropriate latent codes to labels—has no basis for producing meaningful predictions from these out-of-distribution representations. Addressing this limitation will require either adapter layers [Houlsby et al., 2019] that project domain-specific embeddings into the encoder's learned space, or continual learning approaches that allow the encoder to expand its representational capacity without catastrophic forgetting of source-domain knowledge.

6 Practical Implications

6.1 When to Use Each LPF Variant

The results presented in Sections 3 and 4 support clear recommendations for practitioners selecting between LPF-SPN and LPF-Learned in deployment.

LPF-SPN is the appropriate choice for in-domain inference, where its 97.8% accuracy and 0.014 ECE reflect its formal calibration guarantees [Alege, 2026b]. It is particularly suited to scenarios in which interpretability of the aggregation process is a regulatory or audit requirement, since the SPN's explicit factor potentials allow the contribution of individual evidence items to be inspected directly. It is also preferable in small-data regimes where no training data is available for the aggregation networks, since it requires no learning beyond the source-domain VAE training. *However*, LPF-SPN should be avoided for zero-shot transfer to novel domains, where it achieves only 29.0% average accuracy and exhibits catastrophic miscalibration (ECE 0.429); any deployment scenario involving domain shift should default to LPF-Learned.

LPF-Learned is the appropriate choice for cross-domain transfer, achieving 42.5% average zero-shot accuracy and ECE 0.207. Its adaptive evidence weighting generalises more robustly under distribution shift, and its latent-space aggregation prevents the calibration collapse that afflicts LPF-SPN under mismatched factor potentials. The primary caveat is that LPF-Learned requires sufficient source-domain training data for its aggregation networks, and its evidence weights are less directly interpretable than SPN factor potentials. Potential overfitting under few-shot adaptation—as observed in the academic and healthcare domains—also warrants careful monitoring.

6.2 Deployment Recommendations

We propose three operationally distinct deployment strategies based on the available target-domain resources.

Strategy 1: Hybrid Approach. Begin with LPF-Learned for zero-shot transfer to establish a performance baseline. Collect 100–500 labeled samples from the target domain, then apply few-shot adaptation with early stopping conditioned on validation ECE. Monitor calibration metrics throughout; if ECE degrades following adaptation (as in the academic domain), revert to the zero-shot model.

Strategy 2: Domain Similarity Filtering. Prior to deployment, assess the embedding distance d between the target domain's evidence and the source-domain training distribution. If $d < 5.0$, zero-shot LPF-Learned is viable and 100–200 labeled examples are sufficient for meaningful calibration improvement. If $d \geq 5.0$ (as in the legal domain, where $d = 8.7$), zero-shot transfer is likely unreliable and a minimum of 200 labeled target-domain examples should be required before deployment.

Strategy 3: Ensemble Calibration. In scenarios where both variants can be evaluated, weight their predictions by inverse ECE ($w_{\text{method}} \propto 1/\text{ECE}_{\text{method}}$) to combine the structural guarantees of LPF-SPN with the adaptability of LPF-Learned. This approach provides robustness against method-specific failure modes and is particularly appropriate when the domain similarity of a new deployment target is unknown.

6.3 Sample Size Guidelines

Based on our experiments and projections, Table 5 provides target-accuracy-to-sample-count guidance for practitioners planning data collection budgets for new target domains.

Table 5: Sample size guidelines for target-domain adaptation with LPF-Learned. Expected ECE and training time are indicative; actual values depend on domain similarity.

Target Accuracy	Recommended Samples	Expected ECE	Training Time
45–50%	10–50 shots	0.20–0.30	∼5 min
50–60%	100 shots	0.10–0.20	∼10 min
60–70%	200–500 shots	0.05–0.15	∼30 min
70–80%	1,000+ shots	0.03–0.08	∼2 hours
>90%	Full domain training	<0.03	∼8 hours

The cost-benefit inflection point is at 100 shots: zero-shot transfer is free at 42.5% accuracy and ECE 0.207, while 100-shot adaptation achieves 46.3% accuracy and ECE 0.147 at minimal labelling cost. We recommend 100-shot adaptation as the default starting point for any new deployment domain, with the caveat that calibration should be monitored and zero-shot fall-back considered if ECE worsens after adaptation.

7 Related Work

7.1 Cross-Domain Transfer Learning

The problem of domain adaptation has been studied extensively in the context of discriminative models [Pan and Yang, 2010], yet its intersection with *probabilistic* reasoning systems that must preserve calibrated uncertainty estimates under distribution shift has received comparatively little attention. The dominant paradigm in transfer learning for NLP involves fine-tuning large pre-trained language models [Devlin et al., 2019], which achieves high accuracy in target domains but typically degrades calibration significantly [Ovadia et al., 2019]. Domain-adversarial approaches [Ganin et al., 2016, Tzeng et al., 2017] learn representations that are invariant to the source-target domain distinction, trading interpretability for domain-agnostic features. Meta-learning methods [Finn et al., 2017] aim to learn representations that support rapid adaptation from a small number of examples, achieving competitive few-shot accuracy but without explicit uncertainty quantification.

LPF occupies a distinct position in this landscape. Its structured latent representation and explicit probabilistic outputs allow calibration to be monitored and preserved as a first-class objective during transfer, rather than as a post-hoc concern. The comparison with prior work is summarised in Table 6.

Table 6: Comparison of transfer learning approaches across accuracy, calibration, and interpretability. Ratings are qualitative assessments based on reported results and architectural properties.

Method	Accuracy	Calibration	Interpretability
Fine-tuning [Devlin et al., 2019]	High	Poor	Low
Domain-adversarial [Ganin et al., 2016]	Medium	Medium	Low
Meta-learning [Finn et al., 2017]	Medium	Good	Low
LPF-Learned (this work)	Medium	Good	Medium

7.2 Calibration Under Distribution Shift

A growing body of work has demonstrated that neural network confidence estimates deteriorate under distribution shift, even when accuracy remains competitive [Ovadia et al., 2019, Minderer et al., 2021]. Temperature scaling [Guo et al., 2017], while effective for in-distribution calibration, does not generalise to shifted distributions without target-domain validation data. Deep ensembles [Lakshminarayanan et al., 2017] provide more robust uncertainty estimates but incur substantial inference overhead.

Our findings present a nuanced extension of this literature. We show that the choice of aggregation mechanism—rather than the choice of base model or post-hoc calibration technique—is the primary determinant of calibration quality under distribution shift for multi-evidence systems. Specifically, LPF-SPN's multiplicative product inference leads to worse calibration under shift than LPF-Learned's additive latent-space aggregation, despite both variants sharing identical source-domain calibration. This suggests that the architectural inductive biases of multi-evidence aggregation have not previously been considered as a calibration design dimension, and warrants further study.

7.3 Few-Shot Learning for Structured Prediction

Meta-learning approaches for few-shot classification, such as prototypical networks [Snell et al., 2017], typically report 48–60% accuracy on standard benchmarks using 100 training examples. Our LPF 100-shot results (46.3% average) are broadly comparable—with the important distinction that LPF additionally provides calibrated probability distributions over outcomes rather than point class predictions, enabling downstream uncertainty-aware decision-making that purely discriminative few-shot methods cannot support. The non-monotonic learning dynamics we observe at 100 shots connect to observations in the few-shot literature [Snell et al., 2017] regarding the instability of adaptation in the low-data regime, and suggest that regularisation strategies developed for prototypical networks may also benefit LPF few-shot adaptation.

8 Limitations and Future Work

8.1 Current Limitations

Limited domain coverage. The present evaluation spans six target domains, which, while diverse, may not capture the full breadth of semantic shifts encountered in real-world deployment. All six domains share a basic structure of multi-label classification over evidence represented as natural language text, and it remains unclear how the findings generalise to domains with fundamentally different modalities (e.g., time-series data or structured tables) or label spaces of substantially different sizes. Future work should expand evaluation to at least 20 domains, including multi-modal evidence settings.

Fixed encoder architecture. The VAE encoder is frozen throughout all transfer experiments, which serves the important goal of preserving source-domain representations but simultaneously constrains the degree to which the model can adapt to domains with incompatible evidence structure. The legal domain illustrates this limitation acutely: its 32.6% zero-shot accuracy reflects the encoder's inability to produce useful latent representations for legal evidence. Adapter layers inserted between the frozen encoder and the downstream networks [Houlsby et al., 2019] represent a promising approach to domain-specific representation alignment without catastrophic forgetting of source-domain knowledge.

Sample efficiency of few-shot adaptation. The 100-shot adaptation procedure achieves only modest average accuracy gains (+3.8%) and exhibits non-monotonic behaviour that limits its reliability as a practical adaptation strategy. Meta-learning objectives that explicitly optimise for rapid in-context adaptation [Finn et al., 2017] could substantially improve sample efficiency, potentially achieving comparable adaptation quality with 20–30 examples rather than 100.

Computational overhead of ensemble methods. The ensemble strategy described in Section 4.1 increases inference time by approximately 3× relative to single-model inference, which may be prohibitive in latency-sensitive applications. Model distillation of the ensemble into a single network, or efficient approximations based on stochastic depth or dropout sampling, represent natural paths toward reducing this overhead.

8.2 Open Questions

Three questions remain unresolved by the present work. **First:** can transfer success be predicted a priori without running the model? The embedding distance measure (Section 5.3) achieves a correlation of $r = -0.81$, which is useful but imperfect. More sophisticated domain divergence measures—such as Fréchet distances in the latent space or classifier-based domain divergence estimators—may provide more reliable pre-deployment assessments. **Second:** what causes few-shot accuracy to decrease in domains where zero-shot performance is already strong? The academic and healthcare regressions observed at 100 shots suggest either overfitting to sample-specific features or a form of negative transfer in which the adaptation disrupts representations that were already well-aligned. A careful study of the gradient dynamics and loss landscape during adaptation could clarify this mechanism. **Third:** how many source domains are required for universal transfer? The present work uses a single source domain (compliance); the hypothesis that 5–10 diverse source domains would yield a substantially more generalisable encoder warrants investigation.

8.3 Future Directions

Several directions offer the most promising paths toward addressing these limitations. Active learning strategies that query the most informative target-domain examples—rather than sampling randomly—could reduce the data requirement for adaptation from 100 examples to 20–30 while achieving comparable calibration improvements. Multi-source training on several jointly-trained source domains could yield a universal encoder that represents evidence from arbitrary domains, substantially improving zero-shot transfer to semantically distant targets. Online adaptation—continuously updating the adaptation layer with incoming target-domain observations and monitoring calibration drift for automatic retraining triggers—would enable robust deployment in non-stationary environments. Finally, a formal theoretical analysis connecting transfer error to existing domain divergence measures [Pan and Yang, 2010] within a PAC-Bayesian framework would complement the empirical findings of this work with principled guarantees.

9 Conclusion

This paper presents the first comprehensive analysis of cross-domain transfer for Latent Posterior Factors, revealing fundamental trade-offs between structured and learned aggregation mechanisms and providing practical guidance for deployment across heterogeneous domains.

9.1 Summary of Findings

Five findings emerge from this investigation. First, LPF-Learned outperforms LPF-SPN in zero-shot transfer by 13.5 percentage points absolute (42.5% vs 29.0%) while simultaneously maintaining superior calibration (ECE: 0.207 vs 0.429). The mechanism is architectural: latent-space aggregation avoids the multiplicative confidence amplification that causes calibration collapse in SPN product inference under distribution shift. Second, calibration quality critically depends on the aggregation mechanism rather than on post-hoc calibration techniques; this is a design consideration that has not previously been surfaced in the context of multi-evidence systems. Third, domain similarity—quantified as embedding distance in the sentence embedding space—strongly predicts transfer success ($r = -0.81$), with the legal domain representing a complete failure case (32.6% $\approx$ random) due to fundamental semantic incompatibility with the compliance source domain. Fourth, 100-shot domain adaptation improves calibration more reliably than accuracy: ECE reduces by 29% on average ($0.207 \rightarrow 0.147$) while accuracy gains are modest and non-monotonic, with regressions observed in healthcare and academic domains. Fifth, sample efficiency follows an approximately logarithmic trend up to 200 examples, with 500 examples projected to achieve roughly 70–75% accuracy—still 23–28 percentage points below in-domain performance, confirming that the distribution gap cannot be fully closed by fine-tuning alone.

9.2 Practical Impact

For practitioners deploying LPF systems across heterogeneous domains, these findings support four concrete recommendations: use LPF-Learned as the default for any deployment involving domain transfer; budget 100–200 labeled target-domain examples as a cost-effective starting point for adaptation; monitor ECE as a primary deployment quality signal and revert to zero-shot models if calibration degrades; and avoid zero-shot deployment to domains with embedding distance $d > 5.0$ from the source domain without a minimum of 200 labeled target-domain examples.

9.3 Broader Implications

Beyond LPF specifically, this work illustrates that the aggregation mechanism of multi-evidence systems is an under-studied dimension of both transfer performance and calibration robustness. The fundamental tension between structural inductive biases (SPN) and learned flexibility (neural aggregation) mirrors broader debates in machine learning regarding the appropriate role of hard vs. soft constraints in model design. Our results suggest that for distribution-shifted deployment, the costs of structural rigidity—in the form of calibration collapse—outweigh its benefits in interpretability and formal guarantees. Hybrid approaches that combine structural constraints with learned adaptability thus represent a productive direction for future work in probabilistic reasoning systems. Cross-domain transfer remains a challenging open problem even for sophisticated probabilistic frameworks; yet the combination of universal evidence encoding, adaptive aggregation, and few-shot adaptation demonstrated here establishes a viable practical path toward domain-agnostic probabilistic reasoning.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback. Special thanks to the open-source community for PyTorch, Sentence-Transformers, and other libraries that made this research possible.

A Experimental Details

All experiments were conducted on CPU-only hardware to ensure full reproducibility without GPU-specific numerical variation. LPF-SPN inference takes approximately 3.3 ms per query; LPF-Learned takes approximately 5.1 ms. All experiments use seed 11111. The VAE encoder architecture is $[384 \to 256 \to 128 \to 64]$ with ReLU activations, LayerNorm, and Dropout(0.2). The decoder architecture is $[96 \to 128 \to 64 \to 3]$ with ReLU activations and LayerNorm. Training on the compliance source domain uses Adam with learning rate 10^{-3} for 100 epochs.

B Domain Statistics

Table 7: Dataset statistics across all evaluation domains.

Domain	Entities	Evidence/Entity	Train	Test	Classes
Compliance	900	5	600	300	3
Healthcare	900	5	630	270	3
Academic	900	5	630	270	3
Construction	900	5	630	270	3
Legal	900	5	630	270	3
Finance	900	5	630	270	3
Materials	900	5	630	270	3

C Complete Results Tables

C.1 LPF-SPN Zero-Shot Results

Table 8: LPF-SPN zero-shot results across all target domains. Conf (μ) and Conf (σ) denote the mean and standard deviation of prediction confidence.

Domain	Accuracy	F1	ECE	NLL	Brier	Conf (μ)	Conf (σ)
Healthcare	0.267	0.140	0.374	1.441	0.300	0.641	0.061
Academic	0.407	0.193	0.572	2.734	0.380	0.979	0.007
Construction	0.281	0.146	0.682	3.627	0.449	0.964	0.019
Legal	0.326	0.164	0.007	1.099	0.222	0.333	0.000
Finance	0.222	0.171	0.448	1.499	0.318	0.667	0.121
Materials	0.237	0.128	0.493	2.310	0.374	0.730	0.066
Average	**0.290**	**0.157**	**0.429**	**2.118**	**0.341**	**0.719**	**0.046**

C.2 LPF-Learned Zero-Shot Results

Table 9: LPF-Learned zero-shot results across all target domains.

Domain	Accuracy	F1	ECE	NLL	Brier	Conf (μ)	Conf (σ)
Healthcare	0.593	0.248	0.253	1.304	0.227	0.821	0.061
Academic	0.474	0.425	0.214	1.987	0.289	0.518	0.053
Construction	0.363	0.178	0.232	1.317	0.266	0.595	0.056
Legal	0.326	0.164	0.007	1.099	0.222	0.333	0.000
Finance	0.415	0.342	0.280	1.514	0.274	0.661	0.101
Materials	0.481	0.217	0.254	1.275	0.238	0.736	0.045
Average	**0.425**	**0.262**	**0.207**	**1.416**	**0.253**	**0.611**	**0.053**

C.3 100-Shot Adapted Results

Table 10: LPF-Learned results after 100-shot domain adaptation.

Domain	Accuracy	F1	ECE	NLL	Brier	Conf (μ)	Conf (σ)
Healthcare	0.489	0.361	0.074	1.012	0.203	0.456	0.080
Academic	0.341	0.260	0.495	1.556	0.321	0.836	0.132
Construction	0.519	0.423	0.059	1.386	0.238	0.568	0.103
Legal	0.326	0.164	0.007	1.099	0.222	0.333	0.000
Finance	0.630	0.642	0.087	0.795	0.162	0.642	0.146
Materials	0.467	0.225	0.150	1.214	0.239	0.616	0.094
Average	**0.462**	**0.346**	**0.145**	**1.177**	**0.231**	**0.575**	**0.093**

D Per-Domain Analysis Details

D.1 Healthcare Domain

The healthcare domain achieves the best zero-shot performance of any target domain under LPF-Learned (59.3%), attributable to strong semantic overlap with the compliance source domain: both involve structured risk-assessment evidence in the form of reports, scores, and historical records. The top-weighted evidence items identified by the learned aggregator reflect clinically coherent signals:

```
Top weighted evidence (learned aggregator):
  E001: "comprehensive patient assessment" (weight: 0.223)
  E002: "consistent diagnostic results"    (weight: 0.217)
  E003: "clear risk stratification"        (weight: 0.209)
```

Following 100-shot adaptation, accuracy falls to 48.9% (-10.4%), while ECE improves dramatically from 0.253 to 0.074 (-70.8% relative). The accuracy regression is consistent with overfitting to the 100 examples, which may over-represent specific evidence patterns that do not generalise across the healthcare test set.

D.2 Legal Domain

The legal domain constitutes a complete transfer failure for both LPF variants. Both methods achieve 32.6% accuracy—marginally above the random baseline of 33.3%—and output near-uniform predictions ($P(\text{plaintiff}) = P(\text{neutral}) = P(\text{defendant}) \approx 0.333$), producing a degenerate ECE of 0.007 through maximum-entropy uncertainty rather than genuine calibration.

Root cause analysis identifies three contributing factors. First, legal reasoning is semantically incompatible with risk assessment: it requires interpretation of precedent, statutory language, and procedural history, none of which feature in compliance evidence. Second, the evidence modalities are incompatible: legal briefs and case documents differ fundamentally from the audit reports, compliance scores, and inspection records that constitute compliance evidence. Third, the feature distribution is misaligned: legal judgements involve categorical, precedent-dependent reasoning, whereas compliance evidence is predominantly numerical.

The implication is clear: the legal domain requires a minimum of 200 labeled target-domain examples before LPF deployment is viable.

D.3 Finance Domain

The finance domain is the standout success case for 100-shot adaptation, achieving a 21.5 percentage-point accuracy gain (41.5% → 63.0%) with a concurrent ECE improvement from 0.280 to 0.087 (−68.9% relative). The success is attributable to structural alignment between financial and compliance evidence: both domains employ numerical risk-assessment features (credit scores, debt ratios, revenue trends) that the encoder has learned to represent effectively during compliance training. Representative examples from the effective 100-shot training set illustrate this alignment:

```
Effective 100-shot examples:
  - "Debt-to-equity ratio: 0.82 (moderate risk)"
  - "Consistent revenue growth over 3 years"
  - "Strong liquidity position (current ratio: 2.1)"
```

E Calibration Deep Dive

E.1 Calibration Metrics

Three calibration metrics are used throughout this paper. **Expected Calibration Error (ECE)** partitions predictions into confidence bins $\mathcal{B}$ and measures the average gap between bin confidence and bin accuracy:

$$\text{ECE} = \sum_{b \in \mathcal{B}} \frac{|b|}{n} \left|\text{acc}(b) - \text{conf}(b)\right| \tag{1}$$

Lower ECE indicates better-calibrated predictions; a value below 0.05 is generally considered well-calibrated. **Negative Log-Likelihood (NLL)** is a proper scoring rule that penalises overconfident incorrect predictions:

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log P(y_i \mid x_i) \tag{2}$$

The uniform distribution over three classes has NLL $= \log 3 \approx 1.099$; values substantially above this indicate overconfident wrong predictions. The **Brier Score** measures the mean squared error

of probabilistic predictions:

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n} \|\hat{p}_i - e_{y_i}\|^2 \tag{3}$$

where e_{y_i} is the one-hot encoding of the true label.

E.2 Why LPF-SPN Miscalibrates Under Shift

The SPN combines factors multiplicatively:

$$P_{\text{SPN}}(y) \propto \prod_{i=1}^{n} \Phi_i(y)^{w_i} \tag{4}$$

Under distribution shift, individual factors Φ_i become miscalibrated because the decoder that produced them was trained on compliance data. The product operation then amplifies these individual miscalibrations exponentially: each factor assigned moderately high probability to the wrong class compounds with all others to produce near-maximal confidence in incorrect predictions. The academic domain provides the canonical example. With three evidence items having factors $\Phi_1(\text{high}) = 0.85$, $\Phi_2(\text{high}) = 0.82$, $\Phi_3(\text{high}) = 0.88$ and weights $w_1 = 0.7$, $w_2 = 0.8$, $w_3 = 0.75$, the combined unnormalised score is $(0.85^{0.7} \times 0.82^{0.8} \times 0.88^{0.75}) \approx 0.85$, yielding after normalisation $P(\text{high}) \approx 0.979$, against a domain accuracy of only 40.7% and an ECE contribution of 0.572.

E.3 Why LPF-Learned Calibrates Better Under Shift

LPF-Learned's superior calibration under distribution shift arises from two mechanisms. First, its quality network uses posterior variance σ^2 as a signal for evidence reliability, down-weighting items whose representations are uncertain under the shift:

$$\text{quality}_i = \text{QualityNet}([\mu_i,\, \sigma_i]) \tag{5}$$

$$w_i = \text{softmax}(\text{WeightNet}([\text{quality}_i,\, \text{consistency}_i])) \tag{6}$$

Second, aggregation in the continuous latent space avoids the multiplicative error amplification of the SPN product. The resulting predictions in the healthcare domain achieve a mean confidence of 0.821 against a 59.3% accuracy rate (ECE 0.253), representing a substantially better confidence-accuracy alignment than LPF-SPN's 0.641 mean confidence against a 26.7% accuracy (ECE 0.374). The evidence quality weights in this domain— $[0.217, 0.209, 0.203, 0.185, 0.223]$—demonstrate meaningful discrimination between evidence items despite the domain shift.

F Implementation Details

F.1 Domain Adaptation Layer

The domain adaptation layer is a residual transformation with a learnable mixing parameter α:

```
class DomainAdaptationLayer(nn.Module):
    def __init__(self, latent_dim=64):
        self.transform = nn.Sequential(
            nn.Linear(64, 64),
```

```
            nn.LayerNorm(64),
            nn.ReLU(),
            nn.Dropout(0.1),
            nn.Linear(64, 64)
        )
        self.alpha = nn.Parameter(torch.tensor(0.5))

    def forward(self, z):
        z_adapted = self.transform(z)
        return self.alpha * z + (1 - self.alpha) * z_adapted
```

The residual connection parameterised by α preserves source-domain knowledge when target-domain adaptation provides limited signal, while allowing the transform to take over when the 100-shot examples provide sufficient evidence for domain-specific alignment. LayerNorm and Dropout(0.1) are included to reduce overfitting.

F.2 Few-Shot Training Protocol

```
# 100-shot training loop
for epoch in range(10):
    for z, label in few_shot_examples:
        z_adapted = adaptation_layer(z)
        probs = decoder(z_adapted, predicate)
        loss = F.cross_entropy(probs, label)
        loss.backward()
        optimizer.step()
```

Training hyperparameters: learning rate 10^{-4} (chosen to minimise catastrophic forgetting), 10 epochs with early stopping if validation ECE increases, Adam optimiser with weight decay 10^{-5}, full batch of 100 examples per update step.

F.3 Ensemble Strategy

```
# Three ensemble components
options1 = QueryOptions(top_k=5,  temperature=0.8)  # Standard
options2 = QueryOptions(top_k=5,  temperature=1.2)  # Diverse
options3 = QueryOptions(top_k=10, temperature=0.8)  # More evidence

# Aggregate distributions
ensemble_dist = (dist1 + dist2 + dist3) / 3
```

The ensemble combines standard inference, diverse inference (higher temperature increases entropy and improves coverage of the output space), and evidence-rich inference (more evidence items at standard temperature). Simple arithmetic averaging reduces prediction variance and tends to improve calibration by combining predictions from complementary configurations. Total inference overhead is approximately $3\times$ relative to single-model inference.

References

Aliyu Agboola Alege Alege. I know what i don't know: Latent posterior factor models for multi-evidence probabilistic reasoning. *arXiv preprint arXiv:2603.15670*, 2026a. URL https://arxiv.org/abs/2603.15670.

Aliyu Agboola Alege Alege. Theoretical foundations of latent posterior factors: Formal guarantees for multi-evidence reasoning. *arXiv preprint arXiv:2603.15674*, 2026b. URL https://arxiv.org/abs/2603.15674.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Victor Lempitsky. Domain-adversarial training of neural networks. volume 17, pages 2096–2030, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *NeurIPS*, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *UAI*, 2011.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.