Signet: MoE Routing Signatures as Domain Fingerprints for Training-Free LoRA Dispatch

Anonymous authors
Paper under double-blind review

Abstract

Multi-adapter serving on a shared Mixture-of-Experts (MoE) backbone requires a dispatch mechanism that identifies the correct LoRA adapter before decoding begins. Existing approaches either impose latency on the critical serving path or require retraining when adapters are updated, coupling dispatch infrastructure to the adapter lifecycle. We present Signet, a training-free framework that repurposes MoE router logits (computed during every forward pass) as domain-discriminative routing signatures, with no gradient updates, auxiliary models, or additional inference passes.

On an 11-domain benchmark spanning semantically similar sub-domains, Signet achieves **90.5% accuracy** (outperforming the best sentence-encoder baseline by **+8.7pp**) at under 2.1% inference overhead. The performance stems from a structural property of MoE expert specialization: routing signature covariance is inherently anisotropic, and Signet exploits this with a shrinkage-based Mahalanobis classifier suited to the few-shot, high-dimensional calibration regime. The framework extends to open-world deployment via a percentile-calibrated out-of-distribution (OOD) gate and a parameter-free adaptive window for mid-session domain shifts, reaching **93.1%** end-to-end accuracy on mixed in-distribution and OOD traffic.

Because router weights are architecturally excluded from standard LoRA target modules, one-time base-model calibration transfers to all fine-tuned variants without recalibration, permanently decoupling dispatch from the adapter lifecycle. These results establish that MoE routing logits are reliable domain-discriminative signals at zero marginal model cost: adapter dispatch becomes a fixed, one-time calibration step compatible with rapidly evolving adapter registries.

1 Introduction

The deployment of large language models increasingly relies on *multi-adapter serving*: a single pre-trained backbone hosts multiple domain-specific LoRA adapters (Hu et al., 2022; Sheng et al., 2024; Chen et al., 2024), activating the appropriate one per request to serve diverse domains (medical, legal, financial, code) without replicating model weights. This setup requires a *dispatch mechanism* that identifies the correct adapter for each incoming request. The constraint is tight: dispatch must complete within the prefill-to-decode scheduling gap, before the first decode token is generated, because any method that lengthens this gap adds head-of-line latency to every served request, making dispatch overhead a first-class serving concern rather than an amortizable preprocessing cost.

Existing dispatch approaches impose a *computational tax* on every served request. *Semantic approaches* (sentence embedding models (Reimers & Gurevych, 2019) and hidden-state pooling) extract rich representations but require a separate model forward pass or intermediate activations that scale with model size; as MoE models grow to tens of billions of parameters, this encoder overhead approaches or exceeds the prefill latency itself, making such methods increasingly impractical at scale. *Lexical approaches* such as TF-IDF (Joachims, 1998) avoid this overhead but fail on semantically similar sub-domains where domain boundaries are not

lexically apparent. Both classes also require retraining or re-fitting whenever adapters are updated, tightly coupling dispatch infrastructure to the adapter lifecycle.

These limitations share a common cause: all existing approaches treat dispatch as computation *separate from* the model forward pass. But in MoE models, inference already produces a signal that reflects domain structure at no additional cost: the per-layer routing logits that select which experts process each token. Because expert specialization emerges from training on domain-specific content (Fedus et al., 2022), the distribution of routing probabilities, aggregated across tokens and layers into a compact (40 KB for a 40-layer model) yet statistically high-dimensional *routing signature*[1], encodes domain identity as a natural byproduct of how the model processes each input, requiring no extra model calls or GPU kernels to produce.

Signet repurposes this intrinsic signal for adapter dispatch. Because router weights are architecturally excluded from standard LoRA target modules, calibration on the base model transfers to all fine-tuned adapter variants without recalibration, transforming dispatch from a trained component that must be versioned with every adapter update into a one-time calibration step and decoupling dispatch infrastructure from the adapter lifecycle entirely. The central technical challenge is that routing signatures reside in a sparse, high-dimensional space (256–12,288D) with small calibration sets ($N = 10$–50), making covariance estimation ill-conditioned. Signet addresses this with Oracle Approximating Shrinkage (OAS) (Chen et al., 2010) combined with PCA dimensionality reduction, achieving **90.5% accuracy on an 11-domain benchmark** (outperforming the best sentence-encoder baseline by **+8.7pp**) across seven MoE architectures without per-architecture tuning.

In practice, a deployed system also encounters inputs from unregistered domains and multi-turn conversations where the active domain may shift mid-session. A percentile-calibrated OOD gate handles the former without requiring OOD data at calibration time; a parameter-free adaptive window handles the latter by extracting context-aware and turn-sliced signatures from a *single* forward pass via per-token logit slicing, resolving domain shifts without extra forward passes or threshold tuning. Figure 1 illustrates the full pipeline.

Contributions. We make the following contributions:

1. **Routing signatures as gradient-free domain fingerprints**: MoE routing logits (already computed during every forward pass) achieve 90.5% accuracy on an 11-domain benchmark ($N = 50$), outperforming the best sentence-encoder baseline by +8.7pp, across seven MoE architectures (256D–12,288D) with no auxiliary model, extra inference pass, or per-architecture tuning.
2. **LoRA invariance enabling one-time calibration**: Router weights are excluded from standard LoRA target modules. Calibration on the base model is therefore valid for all fine-tuned adapter variants with no recalibration, decoupling dispatch from the adapter lifecycle.
3. **OOD-robust end-to-end pipeline**: A percentile-calibrated PCA+cos gate rejects unregistered domains before classification, reaching 93.1% end-to-end accuracy on mixed traffic with no OOD data required at calibration time.
4. **Parameter-free multi-turn routing**: A single-pass adaptive window strategy resolves mid-session domain shifts via per-token logit slicing, with no extra forward pass or threshold tuning.

2 Related Work

2.1 Mixture-of-Experts Routing Analysis

Expert specialization in MoE models has been studied from multiple angles. Fedus et al. (2022) and Lepikhin et al. (2021) observed that MoE routing allocations exhibit structured patterns across different input types. Chi et al. (2022) analyzed cross-lingual expert utilization, finding that routing distributions serve as implicit language detectors, supporting the hypothesis that routing distributions encode domain-level properties beyond language identity. This body of work characterizes routing patterns as a *phenomenon to be observed*;

[1]Throughout this paper, *routing signature* refers to the per-input vector; *domain fingerprint* refers to its role as a domain identifier.

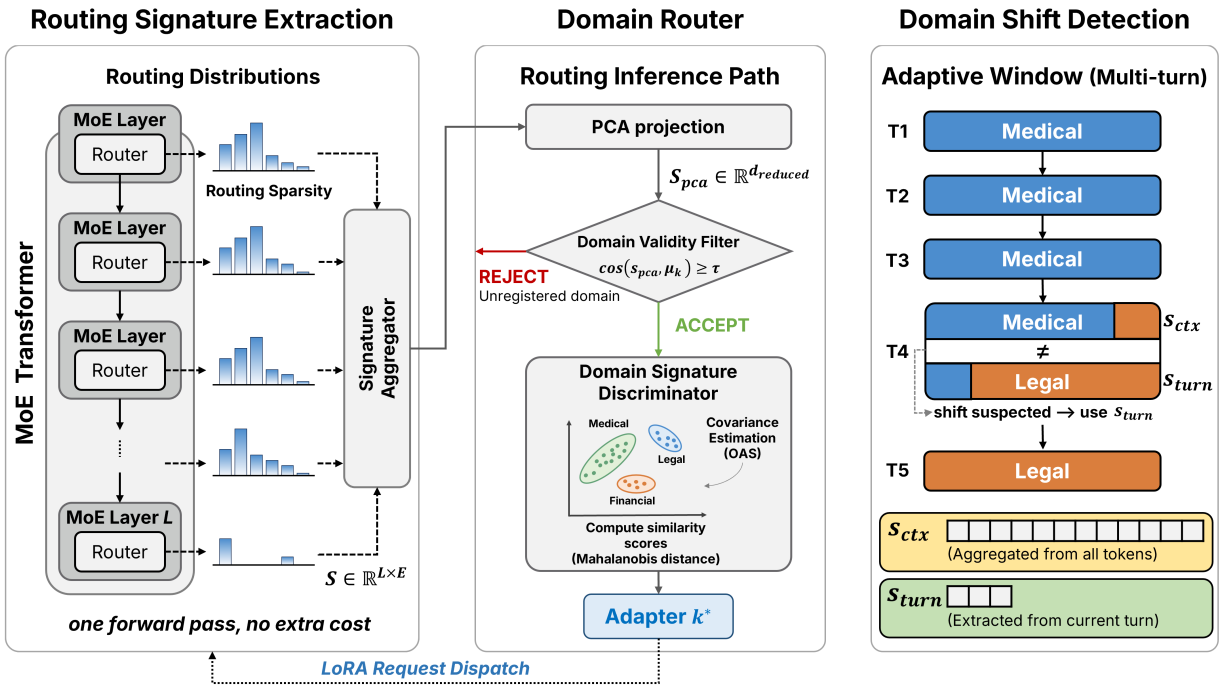

Figure 1: Overview of Signet. **(Left)** Zero-overhead extraction of routing signatures $S \in \mathbb{R}^{L \times E}$ aggregated across all L MoE layers. **(Center)** The Domain Router performs hierarchical inference: a Validity Filter rejects out-of-distribution (OOD) inputs, while a Signature Discriminator selects the optimal adapter k^* via Mahalanobis distance. **(Right)** The adaptive windowing strategy detects mid-session domain shifts (e.g., Medical $\rightarrow$ Legal) by monitoring the discrepancy between history-aggregated (s_{ctx}) and turn-specific (s_{turn}) signatures.

Signet operationalizes the same phenomenon as a *functional dispatch tool*. Recent work confirms that routing statistics are informationally rich: SafeMoE (Kim et al., 2025) and DynMoLE (Li et al., 2025) leverage routing statistics to guide fine-tuning objectives, but exclusively as *training signals* that require a dedicated fine-tuning phase to exploit. Signet repurposes the identical statistics as inference-time byproducts, eliminating the training overhead entirely and extending the utility of routing information downstream of inference to the serving layer.

Most closely related, MoDEM (Simonds et al., 2024) demonstrates that domain-based adapter dispatch is effective, but requires a separately trained BERT classifier and a dedicated encoder forward pass per request. Signet achieves higher accuracy without classifier training or a separate encoder, operating entirely within the existing MoE forward pass.

2.2 Parameter-Efficient Fine-Tuning and Multi-Adapter Serving

LoRA (Hu et al., 2022) and its variants (Dettmers et al., 2023; Zhang et al., 2023) enable efficient domain adaptation through low-rank weight updates; these are the target adapters in Signet's dispatch setting. Multi-adapter serving systems (S-LoRA (Sheng et al., 2024), Punica (Chen et al., 2024), and LoRAX (Predibase, 2024)) achieve high-throughput multi-tenant deployment through batched LoRA execution and memory management. These systems constitute a highly optimized execution engine for running LoRA adapters, but treat adapter selection as a solved prerequisite: they require the upstream dispatch decision to arrive pre-computed and provide no mechanism to produce it. The intelligence layer that maps an incoming request to the correct adapter (without which the execution engine cannot serve any request correctly) is precisely what is missing, and what Signet provides. A separate line of work on adapter *composition* (LoRAHub; Huang et al. 2024) addresses the orthogonal problem of blending adapters for task generalization rather than

routing incoming requests to a fixed domain adapter; it is not directly applicable to the serving dispatch setting.

2.3 Multi-Turn Context and Topic Shift

Topic shift detection in dialogue (identifying when the subject of a conversation changes) is a related line of work in NLP. Lin et al. (2023) use multi-granularity prompting for supervised shift detection; Hwang et al. (2024) show that even large language models struggle with topic shifts without dedicated training data. Both approaches are ill-suited for adapter-serving deployments: they require annotated shift corpora unavailable in production settings, and they output coherence labels or topic-boundary scores rather than actionable domain identities. The fundamental mismatch is one of output: existing topic shift methods answer *"did the topic change?"* (binary coherence); adapter dispatch requires *"which domain is active now?"* (actionable identity). The former is useful for dialogue management; the latter is the necessary and sufficient condition for selecting an adapter. Signet's adaptive window sidesteps these requirements by detecting domain shifts directly from per-token routing logits within the standard MoE forward pass, requiring no labeled corpora and producing adapter identities rather than boundary scores.

2.4 OOD Detection and Covariance-Based Methods

OOD detection for neural networks has been extensively studied (Hendrycks & Gimpel, 2017; Liang et al., 2018). The Mahalanobis distance approach of Lee et al. (2018) fits class-conditional Gaussians on backbone hidden states for OOD scoring; this is architecturally influential. Signet differs in two respects. First, Lee et al.'s approach was designed for the large-data regime, which provides well-conditioned covariance estimates from many labeled samples per class. In the adapter deployment setting, calibration sets are necessarily small ($N = 10$–50 per domain), placing estimation in a regime where $D \gg KN$ and the standard sample covariance is rank-deficient; Signet addresses this with Oracle Approximating Shrinkage (Chen et al., 2010), which provides an analytically optimal shrinkage coefficient without held-out validation data. Second, Signet applies Mahalanobis distance to routing signatures rather than hidden states (a structurally different feature space), yielding a unified estimator for both domain classification and OOD rejection from a single representational space.

3 Method

Signet is designed around four principles that together enable calibration-only, deployment-oriented dispatch:

- **Architectural Invariance**: router weights are excluded from LoRA target modules, so one-time calibration on the base model covers all adapter variants.
- **Statistical Robustness**: PCA combined with OAS covariance shrinkage yields stable classification in the $N \ll D$ few-shot regime.
- **Negligible Overhead**: routing logits are read from already-computed MoE gating operations via forward hooks, with no extra model or inference pass.
- **Open-World Reliability**: a percentile-calibrated OOD gate rejects unregistered inputs before adapter selection, preventing silent misrouting.

3.1 Problem Formulation

Let $\mathcal{F}$ be a pre-trained MoE backbone with L MoE layers, each containing E experts and a learned router $\mathbf{W}_r^{(\ell)} \in \mathbb{R}^{d \times E}$. Let $\{\mathcal{A}_k\}_{k=1}^{K}$ be domain-specific LoRA adapters. Throughout this section, the subscript k denotes a domain index ($k \in \{1, \ldots, K\}$); d_{PCA} (default 200) denotes the PCA-projected dimension and is a distinct symbol. Given input $\mathbf{x}$, the *multi-adapter dispatch problem* is to identify:

$$k^* = \arg\max_k \; P(\text{domain}_k \mid \mathbf{x})$$

using only information available during the standard single-pass MoE forward. We additionally consider the open-world setting where inputs may not belong to any registered domain; here the system must *reject* OOD inputs rather than silently misrouting them. A critical enabling property is that router weights $\{\mathbf{W}_r^{(\ell)}\}$ are architecturally excluded from standard LoRA target modules, making routing signatures numerically invariant to adapter fine-tuning; calibration is therefore performed once on the base model and reused for all subsequently fine-tuned adapters.

3.2 Routing Signature Extraction

For input $\mathbf{x}$, let $m_t \in \{0, 1\}$ be a content mask (1 for content tokens, 0 for padding and special tokens). The router at layer ℓ produces logits $\mathbf{z}_t^{(\ell)} = \mathbf{h}_t^{(\ell)}\mathbf{W}_r^{(\ell)} \in \mathbb{R}^E$ for each token t. We define:

$$\mathbf{p}_t^{(\ell)} = \mathrm{softmax}(\mathbf{z}_t^{(\ell)}), \qquad \mathbf{s}^{(\ell)} = \frac{1}{\sum_t m_t} \sum_t m_t \, \mathbf{p}_t^{(\ell)} \in \mathbb{R}^E \tag{1}$$

In the default setting $m_t = 1$ for all non-padding content tokens, reducing Equation 1 to the uniform token mean. In chat deployments where system-prompt tokens may inflate or distort the token-level average, the mask is adjusted to handle length imbalance between system and user turns; this extension is described in the Multi-Turn Adaptive Window section.

The full routing signature is the concatenation across all MoE layers:

$$\mathbf{s}(\mathbf{x}) = [\mathbf{s}^{(1)}; \mathbf{s}^{(2)}; \ldots; \mathbf{s}^{(L)}] \in \mathbb{R}^D, \quad D = L \times E$$

This signature is extracted via lightweight PyTorch forward hooks on the router modules ($\mathbf{W}_r^{(\ell)}$ for each MoE layer ℓ). E denotes routed experts only; architectures with shared experts (e.g., DeepSeekMoE) contribute no signature dimensions from shared experts. SignetHook hooks the competitive router only.[2] Padding and special tokens are masked during aggregation; only content tokens contribute to $\mathbf{s}^{(\ell)}$. Inputs shorter than 10 content tokens are excluded to ensure signature stability. Memory footprint: $L \times E$ float32 values (5.6 KB for Qwen1.5-MoE-A2.7B; 40 KB for Qwen3.5-35B-A3B).

Algorithm 1: Routing Signature Extraction

Input: MoE backbone $\mathcal{F}$ with L layers and E experts, input sequence $\mathbf{x}$, content mask $m_t \in \{0, 1\}$ (default: 1 for all content tokens), *Stability Threshold* $\tau_{\text{stab}} = 10$ (minimum content token count; inputs with $\sum_t m_t < \tau_{\text{stab}}$ are routed to OOD fallback)

Output: Routing signature $\mathbf{s}(\mathbf{x}) \in \mathbb{R}^D$, $D = L \times E$

Perform a forward pass of $\mathcal{F}(\mathbf{x})$ in *prefill* mode;

Capture router logits $\mathbf{z}_t^{(\ell)} \in \mathbb{R}^E$ for all layers $\ell \in \{1, \ldots, L\}$ and tokens $t \in \{1, \ldots, T\}$;

for *each MoE layer* $\ell = 1, \ldots, L$ **do**

  Compute routing probabilities: $\mathbf{p}_t^{(\ell)} = \mathrm{softmax}(\mathbf{z}_t^{(\ell)})$;

  Compute domain fingerprint via masked temporal averaging:
  $\mathbf{s}^{(\ell)} \leftarrow \frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \cdot \mathbf{p}_t^{(\ell)}$;

end

return $\mathbf{s}(\mathbf{x}) = [\mathbf{s}^{(1)}; \mathbf{s}^{(2)}; \ldots; \mathbf{s}^{(L)}]$; `// Concatenated layer-wise experts`

Signatures are computed from the **prefill phase only**. Decoding-phase logits are conditioned on both the input and already-generated tokens, making them unreliable as pure domain indicators; and adapter selection must precede the first decode step regardless. The hook is invoked once per MoE forward pass and executes in parallel across the batch, so per-request cost does not grow with batch size.

[2] E.g., DeepSeek-V2-Lite contributes $26 \times 64 = 1{,}664$D, not $26 \times 66 = 1{,}716$D. Shared experts are excluded by design: they are activated for every token regardless of input, so their utilization carries no input-conditional routing signal and would only add constant noise to the signature.

Implementation note: on-device accumulation. A pre-allocated GPU buffer accumulates per-token softmax probabilities in-place across all L layers during the single forward pass. Rather than issuing one device-to-host copy per layer, **a single DMA transfer** flushes the entire $L \times E$ buffer at once (40 KB for Qwen3.5-35B-A3B, 5.6 KB for Qwen1.5-MoE-A2.7B), reducing device-to-host transfer overhead by $\sim L\times$ over the naive per-layer baseline (e.g., $40\times$ for Qwen3.5-35B-A3B with $L = 40$ layers). The signature is ready for downstream classification before the first decode step.

3.3 High-Dimensional Covariance Shrinkage for Domain Classification

Prototype construction. Given calibration samples $\mathcal{D}_k = \{\mathbf{x}_i^{(k)}\}_{i=1}^N$ for domain k, the domain prototype is:

$$\boldsymbol{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}(\mathbf{x}_i^{(k)})$$

The high-dimensionality problem. The routing signature dimension $D = L \times E$ ranges from 256 (Mixtral-8x7B) to 12,288 (Qwen3.5-122B-A10B). With $N \ll D$ calibration samples, the sample covariance matrix $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{D\times D}$ is rank-deficient (rank at most $KN - K$). Direct inversion for Mahalanobis distance is numerically unstable, and even pseudo-inverse solutions discard the regularization structure that prevents overfitting. We address this in two steps: first reduce dimensionality via PCA, then estimate a well-conditioned covariance via OAS shrinkage.

PCA pre-projection. To handle the extreme case where $D \gg KN$ (e.g., 10,240D with $K \times N = 550$ samples), we first project signatures to the top $d_{\text{PCA}} = 200$ principal components of the centered signature matrix:

$$\mathbf{s}_{\text{PCA}}(\mathbf{x}) = (\mathbf{s}(\mathbf{x}) - \boldsymbol{\mu}_{\text{global}})\, \mathbf{V}_{200}^{\top}$$

where $\mathbf{V}_{200} \in \mathbb{R}^{200\times D}$ contains the top 200 right singular vectors. This concentrates the domain-discriminative variance (routing patterns are approximately low-rank, as evidenced by the accuracy plateau at $d_{\text{PCA}} = 150$–200 in Appendix A), makes OAS well-conditioned at all $N \geq 10$, and reduces computation from $O(D^3)$ to $O(200^3)$. The value $d_{\text{PCA}} = 200$ is set strictly below the minimum router dimension across all evaluated architectures (Mixtral-8x7B: 256D), ensuring no truncation artifacts on the lowest-dimensional architecture while still covering the full discriminative plateau. All subsequent covariance estimation and classification operate on the d_{PCA}-dimensional projected signatures $\mathbf{s}_{\text{PCA}}$.

OAS shrinkage. With dimensionality reduced to $d_{\text{PCA}} = 200$, we estimate a single shared covariance $\hat{\boldsymbol{\Sigma}}_{\text{pool}}$ from pooled within-class residuals in the PCA space, then apply OAS shrinkage:

$$\hat{\boldsymbol{\Sigma}}_{\text{OAS}} = (1 - \hat{\rho})\, \hat{\boldsymbol{\Sigma}}_{\text{pool}} + \hat{\rho}\, \frac{\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\text{pool}})}{d_{\text{PCA}}}\, \mathbf{I}$$

where $\hat{\rho}$ is the Oracle Approximating Shrinkage coefficient (Chen et al., 2010) computed analytically as:

$$\hat{\rho} = \min\left(\frac{(1 - 2/d_{\text{PCA}})\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\text{pool}}^2) + \operatorname{tr}^2(\hat{\boldsymbol{\Sigma}}_{\text{pool}})}{(n_{\text{pool}} + 1 - 2/d_{\text{PCA}})\left[\operatorname{tr}(\hat{\boldsymbol{\Sigma}}_{\text{pool}}^2) - \operatorname{tr}^2(\hat{\boldsymbol{\Sigma}}_{\text{pool}})/d_{\text{PCA}}\right]},\ 1 \right)$$

The sample count $n_{\text{pool}} = KN$ is the total number of pooled calibration residuals ($= K$ domains $\times$ N samples each).

Intuitively, as $d_{\text{PCA}}/n_{\text{pool}}$ increases (high-dimensional, few-shot regime), $\hat{\rho} \to 1$, collapsing $\hat{\boldsymbol{\Sigma}}_{\text{OAS}}$ to a scaled identity matrix and recovering cosine-equivalent behavior; conversely, as $n_{\text{pool}} \to \infty$ with fixed d_{PCA}, $\hat{\rho} \to 0$ and the estimate converges to the sample covariance.

Remark: Tied covariance. Using a single shared $\hat{\boldsymbol{\Sigma}}_{\text{pool}}$ is justified on two grounds. (1) *Sample complexity*: per-class OAS at $N = 10$ has rank ≤ 9, severely ill-conditioned; the pooled estimator uses KN residuals (550 at $K = 11$, $N = 50$). (2) *Structural homogeneity*: inter-expert correlation arises from model architecture,

not domain content; domain identity is encoded in $\boldsymbol{\mu}_k$, not $\hat{\boldsymbol{\Sigma}}$. Empirically, tied OAS outperforms per-class OAS by +3.9pp at $N = 10$ and +3.0pp at $N = 50$. Uniform diagonal loading (ℓ_2 regularization) is ill-suited here because it applies a fixed eigenvalue shift, inflating noisy dimensions and compressing informative ones equally; OAS adapts $\hat{\rho}$ from the eigenspectrum, preserving large-eigenvalue domain-signal directions while regularizing the flat tail, with no practitioner-specified λ required.

OAS minimizes the expected Frobenius loss between the estimate and the true covariance under a Gaussian model, without requiring a held-out validation set. The pooled residual matrix $\mathbf{R} \in \mathbb{R}^{(KN)\times d_{\text{PCA}}}$ concentrates within-class variance structure across all K domains, enabling the precision matrix $\hat{\boldsymbol{\Sigma}}_{\text{OAS}}^{-1}$ to capture covariance directions shared across domains.

Classification. The squared Mahalanobis distance from signature $\mathbf{s}$ to prototype $\boldsymbol{\mu}_k$ is defined as:

$$d_M^2(\mathbf{s},\, \boldsymbol{\mu}_k) = (\mathbf{s} - \boldsymbol{\mu}_k)^\top \hat{\boldsymbol{\Sigma}}_{\text{OAS}}^{-1} (\mathbf{s} - \boldsymbol{\mu}_k)$$

The shorthand $\boldsymbol{\Omega} \triangleq \hat{\boldsymbol{\Sigma}}_{\text{OAS}}^{-1} \in \mathbb{R}^{d_{\text{PCA}}\times d_{\text{PCA}}}$ denotes the **precision matrix** (the matrix inverse of the OAS-shrunk covariance estimate). Defining the *Mahalanobis matching score* $\mathcal{S}_M \triangleq -d_M^2$ (negated so that argmax selects the nearest prototype), domain assignment is:

$$k^* = \arg\max_k \; \mathcal{S}_M(\mathbf{s}(\mathbf{x}), \boldsymbol{\mu}_k, \boldsymbol{\Omega}) = \arg\min_k \; (\mathbf{s}(\mathbf{x}) - \boldsymbol{\mu}_k)^\top \boldsymbol{\Omega}\, (\mathbf{s}(\mathbf{x}) - \boldsymbol{\mu}_k)$$

Higher $\mathcal{S}_M$ indicates a closer, more likely domain match. The negation converts a *minimum-distance* problem into a *maximum-score* problem: since $d_M^2 \geq 0$, $\arg\max_k(\mathcal{S}_M)$ and $\arg\min_k(d_M^2)$ are equivalent, enabling a uniform argmax interface across all scoring functions (Mahalanobis and cosine). $\boldsymbol{\Omega}$ is a fixed 200×200 matrix regardless of the raw routing signature dimension $D = L \times E$.

Cosine similarity prototype matching is a special case of Mahalanobis distance with $\boldsymbol{\Omega} \propto \mathbf{I}$ when both query and prototypes are unit-normalised, and serves as our primary baseline.

Calibration set integrity. PCA introduces no label leakage. The basis is fit on pooled, mean-centered signatures without class labels. All classification results are evaluated on a held-out test set disjoint from the calibration data used to fit both $\mathbf{V}_{200}$ and $\boldsymbol{\Omega}$. The primary safeguard against overfitting is the strict train/test split: calibration and test sets are written independently with no overlap. As a secondary indicator, the 5-seed CI at $N = 50$ is ± 0.0pp, consistent with the absence of high-variance fitting artifacts. The one in-sample component (gate threshold) is explicitly corrected via leave-one-out cross-validation.

3.4 Two-Stage Deployment Pipeline with OOD Rejection

Motivation. In production, the set of request domains is open-ended. Without explicit OOD detection, every query is assigned to some adapter k^*, even when no registered adapter covers the input domain, creating a silent failure mode that degrades end-user experience. The two-stage pipeline below jointly optimizes an end-to-end metric that credits both correct routing of in-distribution inputs and correct rejection of OOD inputs.

Stage 1 — OOD Gate. We define the gate score as:

$$g(\mathbf{x}) = \max_{k \in 1..K} \cos(\mathbf{s}(\mathbf{x}), \boldsymbol{\mu}_k)$$

The rejection threshold τ is set at the p-th percentile of leave-one-out (LOO) gate scores on the calibration set. For each training sample $\mathbf{x}_i^{(k)}$, the LOO gate score excludes that sample from the prototype $\boldsymbol{\mu}_k$:

$$g_{\text{LOO}}(\mathbf{x}_i^{(k)}) = \max_j \cos\Big(\mathbf{s}(\mathbf{x}_i^{(k)}),\, \boldsymbol{\mu}_j^{(-i)}\Big)$$

where $\boldsymbol{\mu}_j^{(-i)} = \boldsymbol{\mu}_j$ for $j \neq k$, and $\boldsymbol{\mu}_k^{(-i)}$ is domain k's prototype recomputed excluding sample $\mathbf{x}_i^{(k)}$ from its mean. The percentile p controls gate aggressiveness and can be determined independently of the deployment

OOD fraction r. The LOO scores are computed entirely from in-distribution calibration data, so p directly targets the **false rejection rate (FRR)**, an observable quantity:

$$\mathrm{FRR}(p) \approx p/100$$

By construction, setting the p-th percentile threshold means approximately $p\%$ of in-distribution inputs will fall below it. FRR is deployment-agnostic and fully measurable at calibration time without any OOD data. Higher p increases OOD interception at the cost of more false rejections. No held-out OOD data is required at any stage. Once p is chosen to hit a target FRR, E2E performance for any anticipated OOD fraction r can be computed analytically before deployment, using only calibration statistics. This separation between the FRR control knob p and the deployment OOD fraction r is a key operational advantage: FRR targets can be set and verified at calibration time, independently of how much OOD traffic the deployed system will actually encounter.

Stage 2 — Mahalanobis Classifier. Inputs passing the gate are classified via PCA+OAS Mahalanobis distance. Rejected inputs are flagged for fallback handling (e.g., a general-purpose model or human escalation). The PCA-space prototype $\boldsymbol{\mu}_k^{\mathrm{pca}}$ used in Stage 2 is computed as the mean of projected calibration signatures: $\boldsymbol{\mu}_k^{\mathrm{pca}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{s}(\mathbf{x}_i^{(k)}) - \boldsymbol{\mu}_{\mathrm{global}})\mathbf{V}_{200}^{\top}$. By linearity of projection and mean, this is equivalent to projecting the base-space prototype: $\boldsymbol{\mu}_k^{\mathrm{pca}} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\mathrm{global}})\mathbf{V}_{200}^{\top}$; both computations yield identical results and the latter is used in practice.

End-to-end accuracy. The formal metric over mixed traffic is:

$$\mathrm{E2E} = \frac{\sum_i \mathbb{I}[\text{in-dist}_i \wedge \hat{k}_i = k_i^*] + \sum_i \mathbb{I}[\mathrm{OOD}_i \wedge \text{rejected}_i]}{N_{\mathrm{in}} + N_{\mathrm{OOD}}}$$

This metric penalizes both misrouting in-distribution inputs and failing to reject OOD inputs; it is the primary evaluation metric for the end-to-end evaluation.

Design rationale. Both stages operate on PCA-projected signatures ($\mathbf{s}_{\mathrm{pca}}$), which are computed once per request and reused for gating and classification. The gate (Stage 1) uses cosine similarity in the PCA space, a single dot product requiring no matrix inversion; the classifier (Stage 2) uses Mahalanobis distance in the same space, which requires the precision matrix $\boldsymbol{\Omega}$ but is only invoked for inputs that pass the gate. Table 5 confirms the advantage of operating in PCA space: the PCA+cos gate achieves 93.1% E2E and 4.7% FRR vs. 86.9% E2E and 16.7% FRR for the raw-cosine gate, while both achieve 100% OOD rejection. The PCA+cos gate is therefore the recommended default.[3]

False-positive structure for short and generic inputs. Low-content inputs produce near-uniform expert utilization, yielding low cosine similarity to any prototype and triggering OOD rejection, as expected in the absence of domain context, since generic phrasing lacks the concentrated routing signature of domain-specific content. The FRR analysis above applies to domain-appropriate in-distribution inputs; generic short queries are a separate population whose rejection is expected rather than penalized.

3.5 Multi-Turn Adaptive Window Routing

Chat-format inputs. In chat deployments, system-prompt tokens can dominate the routing average when $T_{\mathrm{sys}} \gg T_{\mathrm{user}}$. Setting $m_t = 0$ for system-prompt tokens (user-turn-only aggregation) eliminates this contamination with negligible overhead; inputs shorter than 10 user tokens after masking are sent to the OOD fallback. When the system prompt itself carries the domain context (e.g., a short user query following a long domain persona), user-only aggregation discards the primary domain signal. An empirically motivated **hybrid aggregation** with no extra forward-pass cost handles this:

$$\mathbf{s}_{\mathrm{hybrid}}^{(\ell)} = \alpha\,\mathbf{s}_{\mathrm{user}}^{(\ell)} + (1-\alpha)\,\mathbf{s}_{\mathrm{full}}^{(\ell)}, \qquad \alpha = \min\left(1,\ \frac{T_{\mathrm{user}}}{T_{\mathrm{sys}}}\right) \in [0,1] \tag{2}$$

[3]For deployments requiring finer-grained confidence control, a two-threshold extension partitions requests into OOD-reject, high-confidence routing, and uncertain-zone tiers; the full algorithm and fallback policy options are detailed in the project repository.

$\alpha = 1$ (when $T_{\text{user}} \geq T_{\text{sys}}$) recovers user-only aggregation; as $T_{\text{sys}} \gg T_{\text{user}}$, $\alpha \to 0$ and the signature is dominated by the system prompt. Validation on instruction-tuned chat models is left to future work.

In multi-turn conversations, the routing input at turn t is the sliding-window context $[\text{user}_{t-1}, \text{asst}_{t-1}, \text{user}_t]$ (prev turn + current turn). This provides turn-to-turn continuity but can mask a mid-session domain shift: if the preceding exchange belongs to the old domain, the old-domain signal dominates the routing signature at the switch turn. The **adaptive window strategy** resolves this without any parameter tuning:

1. Perform a **single** forward pass on $[\text{user}_{t-1}, \text{asst}_{t-1}, \text{user}_t]$, capturing per-token router probabilities.
2. Compute $\mathbf{s}_{\text{ctx}}$ by averaging over *all* tokens (context-aware signature).
3. Compute $\mathbf{s}_{\text{turn}}$ by averaging over u_t tokens only via **token-index slicing** on the already-computed per-token probabilities $\{\mathbf{p}_i^{(\ell)}\}$, with **no additional forward pass** and no additional memory. Both $\mathbf{s}_{\text{ctx}}$ and $\mathbf{s}_{\text{turn}}$ are derived from a single set of router logits.
4. Classify both: $\hat{d}_{\text{ctx}} = \textsc{Classify}(\mathbf{s}_{\text{ctx}})$, $\hat{d}_{\text{turn}} = \textsc{Classify}(\mathbf{s}_{\text{turn}})$.
   If $\hat{d}_{\text{ctx}} = \hat{d}_{\text{turn}}$: no shift detected, use $\hat{d}_{\text{ctx}}$ (context-stabilized).
   If $\hat{d}_{\text{ctx}} \neq \hat{d}_{\text{turn}}$: shift suspected, use $\hat{d}_{\text{turn}}$ (latest signal).

This requires no additional forward pass per turn and no threshold tuning. A post-hoc sweep over a score-margin threshold θ on $\hat{d}_{\text{turn}}$ confirms that $\theta = 0$ (any disagreement triggers) is the robust default: the 10-conversation evaluation pool is too small to reliably tune θ, and the binary disagreement signal is already sufficient for the majority of shift detection. The primary risk of $\theta = 0$ is false-positive switches on very short user turns, where $\mathbf{s}_{\text{turn}}$ is noisier due to fewer averaging tokens; the stability threshold $\tau_{\text{stab}} = 10$ from signature extraction provides a first line of defense by routing such turns to the OOD fallback rather than triggering a domain switch on an unreliable signal.

Algorithm 2: Adaptive Window Routing via Token Slicing

Input: turn t, conversation history $\mathcal{H}_t = [(u_1, a_1), \ldots, u_t]$, classifier parameters θ
Output: Domain prediction $\hat{d}_t$

if $t = 1$ **then**
  $\mathbf{x} \leftarrow [u_1]$;
  $n_{\text{pfx}} \leftarrow 0$;
end
else
  $\mathbf{x} \leftarrow [u_{t-1}, a_{t-1}, u_t]$;
  $n_{\text{pfx}} \leftarrow |\text{Tokenize}([u_{t-1}, a_{t-1}])|$;
end

Perform a **single** forward pass on $\mathcal{F}(\mathbf{x})$, capturing router probabilities $\{\mathbf{p}_i^{(\ell)}\}$ for all tokens i and layers ℓ;

$\mathbf{s}_{\text{ctx}} \leftarrow \textsc{MaskedMean}(\{\mathbf{p}_i^{(\ell)}\}, m_i = 1)$; `// Full context signal`
$\mathbf{s}_{\text{turn}} \leftarrow \textsc{MaskedMean}(\{\mathbf{p}_i^{(\ell)}\}, m_i = \mathbb{1}[i > n_{\text{pfx}}])$; `// Current turn signal`

$\hat{d}_{\text{ctx}} \leftarrow \textsc{Classify}(\mathbf{s}_{\text{ctx}}, \theta)$;
$\hat{d}_{\text{turn}} \leftarrow \textsc{Classify}(\mathbf{s}_{\text{turn}}, \theta)$;
if $\hat{d}_{ctx} = \hat{d}_{turn}$ **then**
  return $\hat{d}_{\text{ctx}}$; `// Agreement: follow context`
else
  return $\hat{d}_{\text{turn}}$; `// Disagreement: pivot to latest turn`
end

MaskedMean: layer-wise masked aggregation from Algorithm 1, applied to stored per-token probabilities without an additional forward pass.

3.6 LoRA Invariance and One-Time Calibration

Architectural guarantee. Standard LoRA and QLoRA apply low-rank updates exclusively to attention projections ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$) and/or MLP projections ($\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{up}}, \mathbf{W}_{\text{down}}$). Router weights $\{\mathbf{W}_r^{(\ell)}\}_{\ell=1}^{L}$ precede the expert computation in the MoE layer and are never listed among standard LoRA target modules. Because routing decisions are made by $\mathbf{W}_r^{(\ell)}$ operating on the residual stream prior to expert dispatch, and LoRA does not modify $\mathbf{W}_r^{(\ell)}$, the routing distribution $\mathbf{s}(\mathbf{x})$ is numerically near-invariant to any LoRA fine-tuning that follows the standard target-module convention. LoRA modifies the residual stream entering $\mathbf{W}_r^{(\ell)}$, but the resulting perturbation is empirically negligible (Appendix B).

Practical implication. Calibration (collecting N signatures per domain and computing prototypes, PCA basis, and OAS precision matrix) is performed once on the base model. All subsequently fine-tuned adapters share this calibration without recalibration overhead, because adapter weights never touch the router.

Empirical verification. Routing signatures from the base model and 5 fine-tuned LoRA variants were compared across 13 (seed, checkpoint) pairs on Qwen1.5-MoE-A2.7B; mean cosine similarity was **0.99994** (range: 0.99986–0.99997), confirming numerical near-invariance in practice. Full details are in Appendix B.

4 Experiments

4.1 Models

Seven publicly available MoE architectures are evaluated (Table 1), selected to jointly vary three architectural axes: **expert granularity** ($E = 8$–256 routed experts per layer), **routing depth** ($L = 24$–48 MoE layers), and **signature dimensionality** ($D = 256$–12,288). Of the seven, four belong to the Qwen family; this reflects Qwen's dominant share of publicly available open MoE checkpoints in the 1–15B active-parameter tier at the time of evaluation, and the three non-Qwen models (Mixtral, DeepSeek-V2-Lite, Gemma-4) ensure coverage of distinct routing designs, expert granularities, and training lineages. The set spans both sparse-MoE (Mixtral, Gemma) and fine-grained MoE (DeepSeek, Qwen) routing designs, enabling us to isolate how each architectural axis affects routing signature discriminability. The selection rationale is as follows: Mixtral-8x7B establishes a sparse-MoE baseline at the minimum signature dimension (256D, $E = 8$); Qwen1.5-MoE-A2.7B and DeepSeek-V2-Lite represent lightweight fine-grained architectures with distinct routing depths; Gemma-4-26B-A4B-IT adds a third model family with a separate training lineage; Qwen3-30B-A3B-Thinking tests routing behavior in a chain-of-thought-optimized model; **Qwen3.5-35B-A3B is designated the primary ablation model**. At $D = 10{,}240$ it occupies the middle of the evaluated dimensionality range, allowing ablations to generalize both to smaller (↓ Mixtral, 256D) and larger (↑ Qwen3.5-122B, 12,288D) architectures, and its fine-grained routing ($E = 256$) is representative of the dominant Qwen3 architecture family; and Qwen3.5-122B-A10B provides maximum-scale validation at 12,288D. Architecture details follow their respective technical reports (Jiang et al., 2024; DeepSeek-AI, 2024; Qwen Team, 2024; 2025; Gemma Team, Google DeepMind, 2026).

Table 1: MoE architectures evaluated in this work. Active Params: parameters activated per forward pass; L: number of MoE layers; E: number of routed experts per MoE layer (shared experts excluded); Sig. Dim $= L \times E$.

Model	**Active Params**	L	E	**Sig. Dim**
Mixtral-8x7B-v0.1	∼13B	32	8	256
Qwen1.5-MoE-A2.7B	∼2.7B	24	60	1,440
DeepSeek-V2-Lite	∼2.4B	26	64	1,664
Gemma-4-26B-A4B-IT	∼4B	30	128	3,840
Qwen3-30B-A3B-Thinking	∼3B	48	128	6,144
Qwen3.5-35B-A3B	∼3.5B	40	256	10,240
Qwen3.5-122B-A10B	∼10B	48	256	12,288

4.2 Evaluation Domains

11-domain benchmark. Five categories spanning common enterprise deployment targets are structured into 11 sub-domains; three paired categories (Medical, Legal, Financial) each contain a general sub-domain plus two specialties to create within-category discrimination challenges, while Technical and General serve as singleton controls. Table 2 shows the full structure and the E2E evaluation split: Medical and Legal are registered (in-distribution) as primary high-stakes targets; Financial is designated **hard-negative OOD** because specialist vocabulary overlaps with both registered categories (health economics with medical, regulatory compliance with legal): terms such as *clinical trial costs*, *reimbursement rates*, and *drug liability* straddle financial and medical texts, while *fiduciary duty*, *regulatory filing*, and *compliance framework* bridge financial and legal texts. This lexical overlap makes Financial the hardest rejection case; Technical and General are easy-negative OOD with no vocabulary overlap with the registered set.

Table 2: 11-domain benchmark structure and E2E evaluation split.

Category	Sub-domains (11-domain benchmark)	E2E role
Medical	Gen. Medical, Cardiology, Oncology	Registered
Legal	Gen. Legal, Criminal Law, Contract Law	Registered
Financial	Gen. Financial, Derivatives, Equity Mkts	OOD
Technical	Python Programming	OOD
General	General Text	OOD

50 texts/domain ($N = 50$), 5 random seeds for calibration set sampling. All texts are manually curated domain-representative sentences (1–3 sentences, 25–80 tokens) in authentic domain register, with no slot-filled templates or LLM-generated texts. Texts were authored directly to match realistic production query lengths and register; existing corpora (e.g., PubMed abstracts, legal filings, financial press releases) were not used, as they introduce formatting artifacts, citation noise, and document-length distributions inconsistent with the query lengths typical in serving environments. **Calibration and test sets are strictly disjoint** (50 texts each), written independently with no overlap (see Limitation 1 for caveats on generalization to non-curated texts).

Multi-turn benchmark. Two scenarios are evaluated, each comprising 10 five-turn conversations (see Appendix F for examples): **Scenario A (steady-state)** fixes the domain across all turns, using ambiguous phrasing at T4–5 to probe false-positive shift detections; **Scenario B (domain shift)** pivots at T4 from one domain to another, with T5 continuing in the new domain. The 10-conversation pool limits statistical power; results are indicative rather than definitive.

4.3 Baselines

Three feature types are compared (TF-IDF, SentBERT, routing signatures), each paired with cosine prototype matching as the default classifier. The baseline design deliberately spans the **lexical-to-semantic spectrum**: TF-IDF (lexical extreme) captures surface vocabulary with zero external model cost; SentBERT (`all-MiniLM-L6-v2`, semantic extreme) captures sentence-level meaning via a dedicated 384D encoder forward pass. `all-MiniLM-L6-v2` is selected as the strongest encoder baseline compatible with the prefill-to-decode latency budget: it represents the practical upper bound for encoder-based dispatch in real serving environments, where larger models (e.g., BGE-large, GTE-large) would add 10–50× more overhead than the scheduling gap permits. Signet targets a **best-of-both-worlds** result, matching or exceeding the semantic precision of SentBERT while operating at the inference overhead of a lexical method (no separate encoder; routing features are a byproduct of the standard serving prefill). Neither external baseline requires access to the served model. For routing signatures, PCA+OAS Mahalanobis distance is additionally evaluated as the proposed classifier:

- **TF-IDF**: scikit-learn `TfidfVectorizer`; vocabulary fitted on calibration set.

- **SentBERT**: `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019); 384D embeddings from a separate encoder forward pass.
- **Routing signature (cosine)**: proposed features with cosine prototype matching.
- **Routing signature (PCA+OAS)**: proposed features with PCA+OAS Mahalanobis distance.

All baselines are training-free: they require only calibration samples, with no separate classifier training step. This ensures a fair comparison within the training-free setting. MoDEM (Simonds et al., 2024) (the only prior work directly comparable in task scope) requires classifier training and is excluded from the primary comparison; its published accuracy serves as a reference for the trained-classifier upper bound, and its architectural trade-offs are discussed in the Related Work and Discussion sections. A logistic regression variant (Routing+LR) is included in Appendix C as an internal learned-classifier ablation; it collapses to chance on routing signatures, confirming that PCA+OAS is the correct comparison point.

For OOD gating, two scoring functions are compared. Full-dim OAS gating is computationally impractical at the signature dimensions evaluated here and is included only as an oracle reference in Table 3. The PCA+cos gate achieves comparable or better performance at a fraction of the cost (Table 5):

- **Cosine gate**: $\max_k \cos(\mathbf{s}(\mathbf{x}), \boldsymbol{\mu}_k)$ as gate score (higher = in-dist).
- **PCA+cos gate**: cosine score computed in the PCA-projected space.

4.4 Metrics and Evaluation Protocol

Computational setup. Domain classification experiments use 4-bit quantization (bitsandbytes NF4) on a single NVIDIA A100 80GB GPU; latency benchmarks use BF16 precision on two NVIDIA A100 80GB GPUs. NF4 quantization does not materially alter routing signatures: router weights are never quantization targets (they are excluded alongside LoRA target modules), so only downstream linear projection weights are affected. This is consistent with the LoRA invariance result (Appendix B): routing signatures are empirically insensitive to weight perturbations in non-router modules, and NF4 quantization constitutes a perturbation of comparable or smaller magnitude. Any residual quantization-induced routing shift would affect all methods equally and cannot explain cross-method accuracy gaps. As a reference bound, the base-to-LoRA cosine similarity measured in Appendix B (≥ 0.99986) indicates the scale of routing perturbations from non-router weight changes; NF4 quantization targets the same weight classes and constitutes a perturbation of comparable or smaller magnitude, strongly suggesting NF4-vs-BF16 cosine similarity in the same range. A direct per-signature NF4-vs-BF16 measurement is left to future work. SignetHook overhead is measured using $N = 50$ paired wall-clock samples of `llm.generate()` at batch sizes 1–32 (`seq_len=128`, `max_tokens=1` to isolate prefill-phase cost). The 40 KB SignetHook buffer (Qwen3.5-35B-A3B) is negligible relative to the KV cache footprint at typical serving batch sizes (KV cache scales at $\sim$1 MB per token per layer in BF16) and does not compete for memory bandwidth or allocation budget with the serving stack.

- **Domain classification**: accuracy on a held-out test set completely disjoint from calibration. All classification results use this standard held-out split. LOO evaluation is not used as a primary metric.
- **OOD detection**: OOD rejection rate (fraction of OOD inputs correctly rejected below the calibrated threshold) and false rejection rate (FRR). Prototypes and covariance fitted on the calibration set; evaluation on a disjoint held-out set with no overlap.
- **End-to-end accuracy**: fraction of correctly handled requests across 300 in-distribution + 250 OOD inputs (550 total); the 300+250 split follows directly from the 6 in-distribution and 5 OOD domain partition with 50 test samples per domain. Credits correct routing for in-dist and correct rejection for OOD.
- **Multi-turn metrics**: per-turn accuracy (fraction of correctly routed turns at each position) and **MTTS** (Mean Turns to Switch), a system-level *adaptation speed* metric: the number of turns

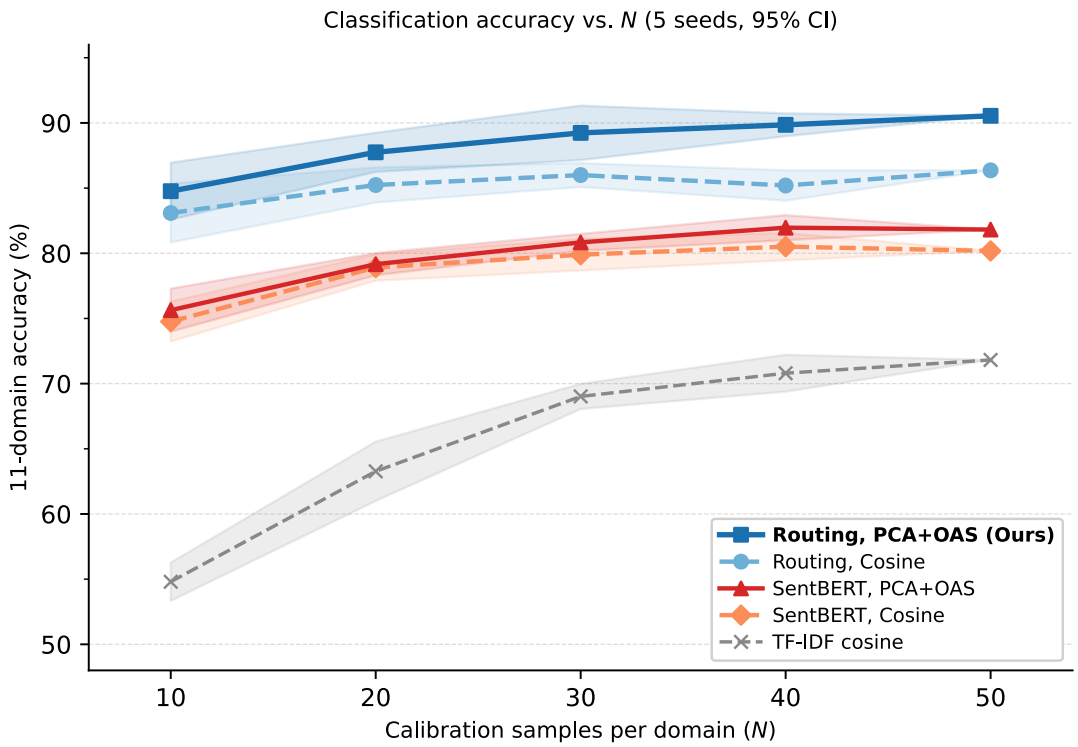

Figure 2: Classification accuracy vs. calibration samples per domain (N), averaged over 5 seeds with 95% CI shaded (Qwen3.5-35B-A3B, 11-domain benchmark). Routing PCA+OAS (Ours) leads all baselines across the full $N = 10$–50 range, with the advantage widening as N increases; routing signatures already surpass SentBERT at $N \approx 15$, illustrating the few-shot advantage of leveraging model-internal features.

from the domain pivot until the first correct domain classification, inclusive. MTTS=1 means the system adapts immediately at the pivot turn; MTTS=2 requires one recovery turn; lower is strictly better. MTTS complements per-turn accuracy by capturing how quickly the system responds to environmental change, rather than just whether it eventually responds correctly.

- **Statistical reporting**: 95% CI via t-distribution ($df = n - 1$), reported as half-width.[4] Pairwise method comparisons use the McNemar test (Bonferroni-corrected), reported alongside the relevant analyses.

5 Results

5.1 Routing Signatures vs. Baselines (11-Domain)

On the 11-domain benchmark (Qwen3.5-35B-A3B), PCA+OAS outperforms routing cosine by **+4.1pp at N=50** and **+1.7pp at N=10** (Table 3), with all pairwise differences significant at $p < 0.001$ (McNemar test, Bonferroni-corrected). This advantage reflects the *anisotropic structure* of routing signatures: expert specialization concentrates domain-discriminative variance into a small number of large-eigenvalue directions that OAS's eigenspectrum-adaptive precision matrix $\mathbf{\Omega}$ is uniquely positioned to exploit; isotropic classifiers that treat all directions equally cannot leverage this structure. At $N = 10$, the narrower gap reflects a covariance estimation constraint (fewer calibration samples to estimate $\mathbf{\Omega}$ reliably); the advantage widens consistently from $N \geq 20$ (Figure 2). Notably, routing signatures already surpass SentBERT at $N \approx 15$: a

[4]Displayed ±0.0% is a rounding artifact; actual CI < 0.05pp.

new domain adapter can be brought into full Signet dispatch with fewer than 20 calibration examples and zero model retraining, directly enabling the rapid adapter onboarding that motivates training-free dispatch. The choice of $d_{\text{PCA}} = 200$ is validated by a full ablation in Appendix A: accuracy plateaus at $d_{\text{PCA}} = 150$–200 (90.5%) and is conservative relative to the minimum architectural dimension (Mixtral: 256D). The end-to-end pipeline evaluation uses a 6/5 holdout split from the same benchmark, with no external OOD corpora.

The 0.7–1.1pp gap between PCA+OAS and full-dim OAS[5] is not statistically significant (McNemar, Bonferroni; TOST equivalence, Appendix C), confirming that PCA projection to $d_{\text{PCA}} = 200$ is not a lossy compression: at $K{\cdot}N = 550$ samples, the 200×200 PCA-space covariance (≈20K free parameters) is reliably estimated, whereas the full 10,240×10,240 covariance (≈52M free parameters) remains underdetermined even at $N = 50$. Comparisons with learned classifiers (SentBERT+LR, routing+LR) are in Appendix C.

The confusion matrices (Figure 3) reveal a key structural contrast: routing signatures confine nearly all errors *within* the correct semantic block (e.g., cardiology↔general_medical), whereas SentBERT exhibits substantial *cross-category* confusion (financial misclassified as legal). From a serving-system perspective, intra-category errors are **structurally safe fallbacks**: misrouting cardiology to general_medical activates an adapter in the same clinical domain, yielding qualitatively similar responses; cross-category errors (e.g., financial→legal) activate a categorically wrong adapter and degrade response quality irreparably. In user-facing applications, intra-category errors produce responses that remain clinically or legally coherent (if not perfectly specialized), whereas cross-category errors may generate authoritative-sounding responses that are factually misaligned with the user's actual domain, a potentially more harmful failure mode in high-stakes deployments. Signet's error structure therefore degrades gracefully even at the cases it mislabels. Within routing signatures, financial sub-domains are hardest at 85.3% (derivatives↔equity_markets most confused), legal 85.3% (criminal/contract), and medical 97.3%, all at $N = 50$.

Table 3: 11-domain classification accuracy (N calibration samples/domain, 5 seeds, 95% CI). Per-domain breakdown in Figure 3.

Method	**N=10**	**N=30**	**N=50**
Lexical baseline			
TF-IDF cosine	54.8% ± 1.5%	69.0% ± 1.0%	71.8% ± 0.0%
Semantic embedding baseline†			
SentBERT cosine	74.8% ± 1.5%	79.9% ± 1.2%	80.2% ± 0.0%
SentBERT OAS (full dim)	74.5% ± 2.4%	80.6% ± 0.4%	81.8% ± 0.0%
SentBERT PCA+OAS	75.6% ± 1.7%	80.8% ± 0.7%	81.8% ± 0.0%
Routing signature (ours)‡			
Routing sig. cosine	83.1% ± 2.3%	86.0% ± 0.9%	86.4% ± 0.0%
Routing sig. OAS (full dim)	84.0%* ± 1.6%	88.1%* ± 1.8%	89.8%* ± 0.0%
Routing sig. PCA+OAS (Ours)	**84.8%* ± 2.2%**	**89.2%* ± 2.1%**	**90.5%* ± 0.0%**

†SentBERT: `all-MiniLM-L6-v2`, 384D. ‡Routing model: Qwen3.5-35B-A3B ($D = 10{,}240$). *$p < 0.001$ vs. routing sig. cosine baseline (McNemar, Bonferroni); * applies within the routing-signature group only.

5.2 Overhead and Latency

SignetHook adds **2.7 ms (+2.07%)** at BS=1, **below 2% of total serving latency** and roughly half the overhead of SentBERT-based dispatch (≈5 ms), while delivering **+8.7 pp** higher accuracy (Table 4). This places Signet squarely within the **prefill-to-decode scheduling gap**, enabling inline dispatch without materially extending first-token latency (TTFT): in vLLM-style serving, the 2.7 ms overhead is absorbed within the prefill scheduling transition, and the <2.1% TTFT increase at BS=1 falls within the scheduling jitter observed across batch sizes. No additional model calls are required: routing features are a byproduct of

[5]Full-dim OAS is included as an oracle reference only.

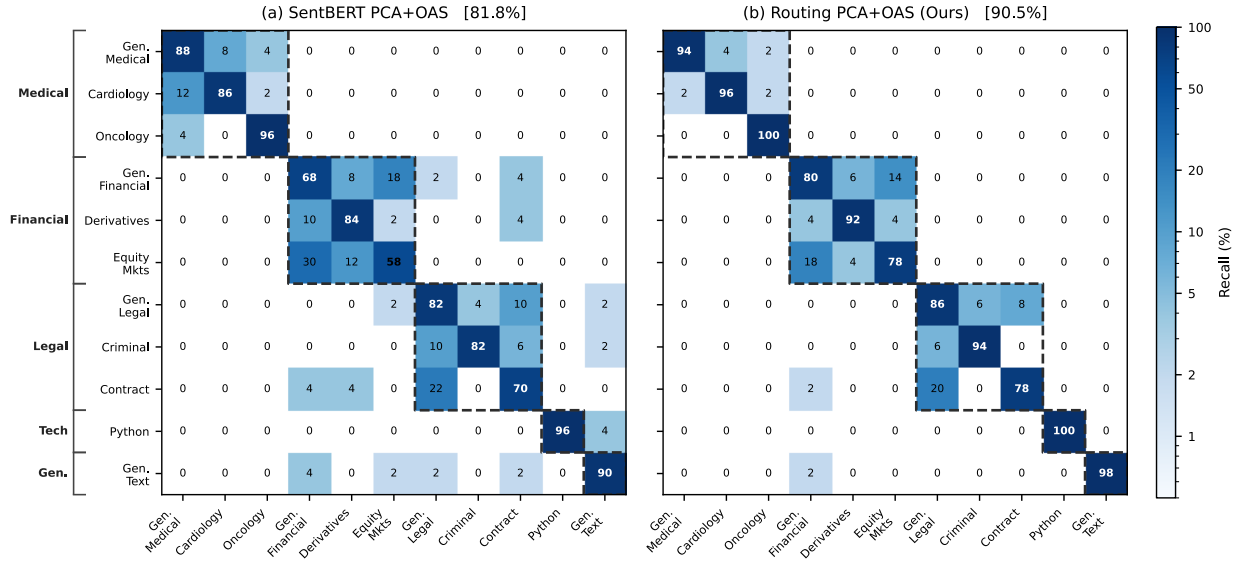

Figure 3: Row-normalised recall confusion matrices on the 11-domain benchmark ($N = 50$, 5 seeds). **(a) SentBERT PCA+OAS** (`all-MiniLM-L6-v2`, 81.8%) vs. **(b) Routing PCA+OAS** (90.5%, ours, Qwen3.5-35B-A3B), the best-performing configuration per method at $N = 50$. Log-scale colormap; true-zero cells are white; dashed boxes delineate the five semantic category blocks; errors *within* a dashed block are **intra-category** (near-miss, structurally safe fallback), while errors *across* blocks are **inter-category** (categorically wrong adapter). Whereas SentBERT exhibits substantial inter-category confusion (e.g., Financial misclassified as Legal), routing signatures confine nearly all errors to intra-category mismatches within the correct semantic block.

the forward pass, held in a 40 KB GPU buffer (Qwen3.5-35B-A3B) flushed via a single GPU→CPU DMA transfer after prefill completes (see Experiments for benchmark configuration).

Table 4: vLLM `generate()` latency with and without SignetHook

BS	Bare (ms)	Hook (ms)	Δ (ms)	Overhead	Bare tok/s	p-val
1	130.1 ± 4.0	132.9 ± 4.7	+2.70	+2.07%	984	<0.001*
4	251.3 ± 7.9	256.1 ± 6.7	+4.60	+1.84%	2,038	<0.001*†
8	254.9 ± 5.9	259.1 ± 7.4	+4.14	+1.62%	4,017	0.008*†
16	248.8 ± 6.3	253.2 ± 6.7	+4.25	+1.71%	8,232	0.001*†
32	252.2 ± 5.8	256.2 ± 5.9	+3.90	+1.55%	16,241	<0.001*

*$p < 0.05$ (Wilcoxon signed-rank). Δ = Hook(ms) − Bare(ms), reported as 10% trimmed mean to reduce sensitivity to scheduling outliers. Bare tok/s: throughput without SignetHook at the corresponding batch size, computed as BS / Bare(ms) × 1000. Measured at `max_tokens=1`; overhead amortizes over longer sequences. †BS=4,8,16: 1 outlier pair each removed from the trimmed-mean calculation (IPC scheduler stalls identified by |diff−median| > 3×IQR).

The dominant cost is the one-time GPU→CPU copy, implemented as a single 40 KB DMA flush rather than $L = 40$ per-layer transfers (a ∼40× reduction in transfer round-trips over a naive per-layer baseline). Overhead is consistent across all batch sizes: 2.70 ms (+2.07%) at BS=1, stabilising at 3.9–4.6 ms (+1.55–1.84%) for BS=4–32. All batch sizes are statistically significant (Wilcoxon $p \leq 0.008$); overhead decreases slightly with batch size as GPU compute increasingly dominates. SignetHook requires zero extra VRAM beyond the forward pass.

Relative to SentBERT dispatch (≈5 ms, measured in the same vLLM serving environment at BS=1 via a sequential `sentence-transformers` encode call), Signet adds only +2.7 ms at BS=1 (roughly 2×–4× lower overhead) while achieving +8.7 pp higher accuracy. The full accuracy-vs.-overhead analysis is in Figure 5, which shows that Signet is the only method where accuracy and overhead are *anti-correlated*: higher-accuracy dispatch incurs lower overhead, because routing features are a byproduct of the existing forward pass rather than an addition to it.

5.3 OOD Detection and End-to-End Pipeline

Two gate configurations are evaluated on mixed traffic ($N_{\text{in}} = 300$, $N_{\text{OOD}} = 250$) using a percentile-calibrated threshold ($p = 20$, targeting ≈20% theoretical FRR) from LOO in-distribution scores, with no OOD data required at deployment (Table 5). Without a gate, naive cosine routing achieves only **49.6% E2E**[6]: adding a raw-cosine gate improves E2E to **86.9%** (+37.3pp) with 100% OOD rejection, but incurs FRR=16.7% (below the theoretical 20%, as the LOO bound is conservative in the raw-cosine space). The PCA+cos gate further reduces FRR to **4.7%**, far below the theoretical bound: PCA projection concentrates domain-discriminative variance into directions that geometrically tighten the in-distribution signature cluster, raising their minimum cosine similarity while pushing OOD signatures further below threshold, without any additional calibration data. This geometric improvement maintains 100% OOD rejection while improving in-distribution routing accuracy (91.6%), yielding **93.1% E2E** (+43.5pp over no-gate). The residual 4.7% false rejections arise from short or generic in-distribution queries that produce near-uniform expert utilization and fall below the gate threshold; as described in the Method section, this is expected behavior in the absence of sufficient domain-specific content, and such queries would benefit from the OOD fallback path regardless of the gate setting.

Table 5: End-to-end accuracy on mixed traffic ($N = 50$, $p = 20$)

Configuration	FRR	OOD Rej.	E2E	In-dist acc†
No gate (cosine only)	0%	0%	49.6%	91.0%
Cosine gate ($p = 20$)	16.7%	**100.0%**	86.9%	91.2%
PCA+cos gate ($p = 20$)	**4.7%**	**100.0%**	**93.1%**	**91.6%**

†Routing accuracy computed only over in-distribution queries that pass the gate (i.e., are not rejected). $N_{\text{total}} = N_{\text{in}} + N_{\text{OOD}} = 300 + 250 = 550$. E2E accuracy credits both correct in-distribution routing and correct OOD rejection.

5.4 Multi-Turn Routing Evaluation

In production, adapter-dispatch requests arrive as multi-turn conversations where the active domain may shift mid-session. Routing is evaluated on the 5 coarse-grained categories and two scenarios described in the experimental setup. Calibration pools 50 texts per category from the `domain_texts_train` calibration pool; the PCA+OAS classifier is unchanged from the classification experiments.

Baseline (window=1). On Scenario A this achieves 100% across all turns. On Scenario B, T1–T3 accuracy is 100% but collapses to **20%** at the switch turn T4, equivalent to random chance for 5 categories; the mechanism is analyzed in the Composable Robustness Stack section.

Adaptive window routing. The adaptive window strategy requires no threshold tuning and no additional forward pass: within the single window=1 forward pass over [user_{t-1}, asst_{t-1}, user_t], two routing signatures are extracted via per-token logit slicing, namely $\mathbf{s}_{\text{ctx}}$ (mean over all tokens) and $\mathbf{s}_{\text{turn}}$ (mean over user_t tokens only, identified by prefix-length counting). Their predictions $\hat{d}_{\text{ctx}}$ and $\hat{d}_{\text{turn}}$ are compared: if they agree, $\hat{d}_{\text{ctx}}$ is used; if they disagree, a domain shift is suspected and $\hat{d}_{\text{turn}}$ is used instead.

[6]E2E = 91% × 300/550 ≈ 49.6%; all 250 OOD inputs are silently misrouted.

Table 6: Multi-turn routing accuracy (Qwen3.5-35B-A3B, 5-domain, PCA+OAS, $N = 50$, 10 conversations per scenario)

Scenario	Method	T1–3	T4 (switch)	T5	Overall	MTTS
A (steady-state)	Baseline (window=1)	100%	100%	100%	100%	—
	Adaptive (ours)	100%	100%	100%	100%	—
B (domain shift)	Baseline (window=1)	100%	20%	100%	84%	1.8
	Adaptive (ours)	**93%**	**80%**	**90%**	**90%**	**1.2**

MTTS = Mean Turns to Switch (turns to first correct classification after domain pivot; ideal = 1). —: MTTS undefined for Scenario A, which contains no domain pivot.

As shown in Table 6, adaptive routing raises T4 accuracy from 20% to **80%** and improves Scenario B overall from 84% to **90%** (+6pp). With MTTS=**1.2**, 8 of 10 conversations correctly reclassify the active domain *at the pivot turn itself*, achieving **zero-shot domain adaptation** with no prior signal from the new domain, no threshold tuning, and no additional forward pass. From an adaptation-speed perspective, MTTS=1.2 means Signet responds to domain shifts within a single conversational exchange in 80% of cases, a critical property for multi-domain dialogue systems where delayed adaptation can cascade into multiple turns of misrouted responses before the user notices the error. The remaining 2 conversations recover at T5, so all conversations converge within one recovery turn of the pivot (MTTS ≤ 2). Baseline MTTS=1.8 indicates that in the 10-conversation pool, some conversations recover at T4 while others require one additional turn (T5), averaging to 1.8 turns from the pivot. We recommend $\theta = 0$ as the robust default: any disagreement triggers the switch without threshold tuning. The best threshold found post-hoc on this pool ($\theta = 108.5$, 96%) is overfit to the 10-conversation evaluation set (see Limitation 1). The three false-positive triggers (two at T3 and one at T5, on semantically borderline queries where the turn-sliced signal incorrectly disagrees with the context-aware signal) cause minor degradation on T3 (80% vs. 100%) and T5 (90% vs. 100%).

5.5 Multi-Architecture Generalization

We evaluate PCA+OAS across all seven MoE architectures to assess generalization beyond the primary Qwen3.5-35B-A3B setting.

Table 7: Multi-architecture classification accuracy on the 11-domain benchmark ($N = 50$/domain, PCA+OAS, $d_{\text{PCA}} = 200$)

Architecture	Sig. Dim	Cosine (%)	PCA+OAS (%, $d_{\text{PCA}} = 200$)	Gain
Mixtral-8x7B-v0.1	256	71.6	80.4	+8.7pp
Qwen1.5-MoE-A2.7B	1,440	74.7	86.4	**+11.6pp**
DeepSeek-V2-Lite	1,664	75.8	83.6	+7.8pp
Gemma-4-26B-A4B-IT	3,840	79.1	85.1	+6.0pp
Qwen3-30B-A3B-Thinking	6,144	79.1	86.0	+6.9pp
Qwen3.5-35B-A3B	**10,240**	**86.4**	**90.5**	**+4.1pp**
Qwen3.5-122B-A10B	12,288	87.8	90.4	+2.5pp

Full d_{PCA} sweep for Qwen3.5-35B-A3B is in Appendix A.

PCA+OAS consistently outperforms cosine on every architecture, with gains ranging from +2.5pp (Qwen3.5-122B-A10B) to **+11.6pp** (Qwen1.5-MoE-A2.7B), robust across the full 256→12,288D range, across both sparse-MoE (Mixtral, Gemma) and fine-grained MoE (DeepSeek, Qwen) routing designs, confirming **architecture-agnostic generalization**. The dramatic +11.6pp gain on Qwen1.5-MoE (1,440D) is not accidental: at low dimensionality, the $N \ll D$ regime is most severe and OAS covariance shrinkage provides maximum corrective benefit over the rank-deficient sample covariance. Specifically, at $D = 1{,}440$ with

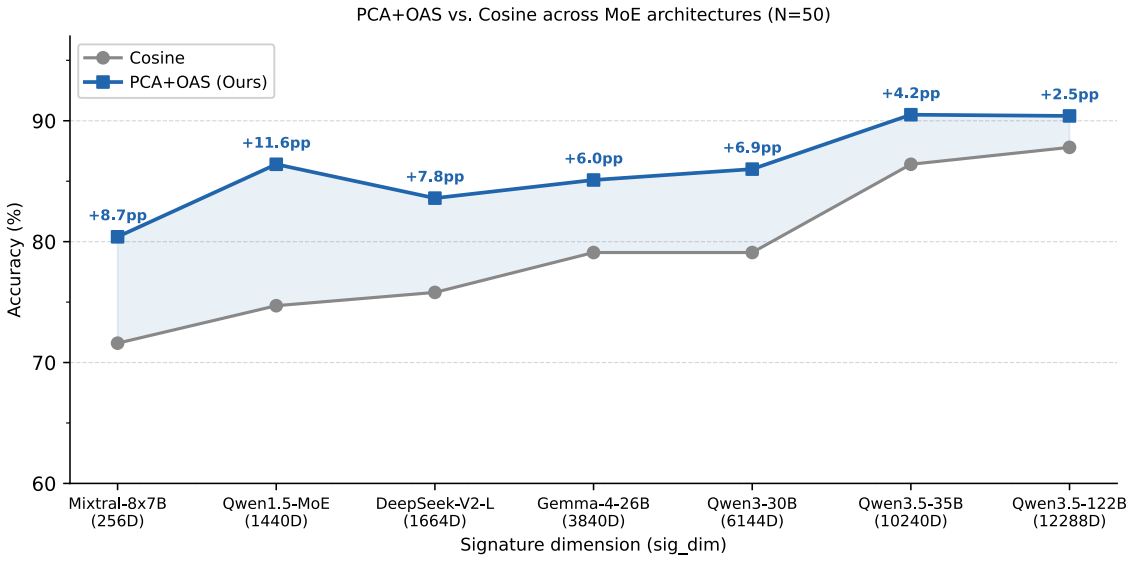

Figure 4: PCA+OAS vs. cosine accuracy across seven MoE architectures (256D–12,288D, $N = 50$, $d_{\text{PCA}} = 200$). Gains range from +2.5pp (Qwen3.5-122B-A10B) to +11.6pp (Qwen1.5-MoE-A2.7B); the shaded region shows the PCA+OAS advantage at each architecture. Gains are largest for lower-dimensional architectures (Qwen1.5-MoE-A2.7B, DeepSeek-V2-Lite), where the $N \ll D$ regime is most severe and covariance shrinkage has the most structure to exploit.

$K \times N = 550$ pooled residuals the sample covariance rank is severely deficient and $\hat{\rho}$ approaches its maximum, concentrating all regularization onto the few large-eigenvalue inter-domain axes. The +11.6pp result is therefore direct empirical validation of the OAS shrinkage mechanism: the gain is largest precisely where the covariance estimation problem is hardest. Cosine accuracy is broadly positively correlated with dimensionality (71.6%→87.8%; Figure 4), though model size, depth, and expert count co-vary, so this should not be read as a scaling law. The 0.1pp gap between Qwen3.5-35B (10,240D, 90.5%) and Qwen3.5-122B (12,288D, 90.4%) confirms the *dimensionality saturation* predicted by the intrinsic-rank bound: once D exceeds the $K - 1$ inter-domain axes, additional dimensions contribute no discriminative information.

LoRA invariance is confirmed empirically in Appendix B: one-time calibration on the base model transfers to all fine-tuned adapter variants without recalibration.

6 Discussion

MoE expert specialization simultaneously produces the domain signal Signet needs to dispatch adapters *and* the statistical structure that makes it classifiable without supervision. Three interlocking properties explain the empirical results: *anisotropic covariance* arising from expert specialization determines which classifier families succeed or fail; *negligible overhead* because the signal is a byproduct of the forward pass; and *parametric simplicity* because OOD detection, domain-shift tracking, and short-query robustness all repurpose the same signal without labeled data.

6.1 Structural Basis for Routing-Signature Classification

Routing signatures occupy a distinctive statistical regime: high dimensionality (256–12,288D), small calibration sets ($N \approx 10$–50), and structured covariance from MoE expert specialization. Domain-specialized experts activate at high probability for their target domain and near-zero for others, producing a covariance matrix with a small number of large eigenvalues (inter-domain axes) and many small ones (within-domain noise). This anisotropy is not a training artifact but an inevitable consequence of the MoE inductive bias: load-balancing

pressure forces each expert to specialize on a subset of the input distribution, and that specialization is precisely what concentrates domain variance into a small number of large-eigenvalue axes. The routing logits are therefore an *unsupervised expression* of domain identity that the model constructs during pretraining as an organizational response to data diversity, not as a dispatch signal. The differential gain makes this concrete: PCA+OAS adds up to +11.6pp on routing signatures but only +1.6pp on SentBERT embeddings where contrastive training has already smoothed the covariance toward isotropy (Tables 3, 7). The discriminative eigenspectrum is bounded by at most $K - 1$ inter-domain axes, independent of raw signature dimension D; the dimensionality saturation observed across architectures (two architectures separated by 2,048D achieving essentially identical accuracy) follows directly from this intrinsic-rank ceiling. The saturation pattern implies that routing logits, across architectures of vastly different scale, project input domain onto a shared low-dimensional manifold of at most $K - 1$ axes: the model's internal domain map is not proportional in complexity to its parameter count or signature dimension, but to the number of domains encountered during pretraining. This low-dimensional structure is a direct consequence of MoE specialization (not a coincidence) and explains why routing signatures remain discriminative even at the smallest evaluated dimensionality (Mixtral: 256D).

This anisotropy determines which classifier families succeed. OAS identifies which dimensions carry inter-domain signal and which carry within-domain noise, then applies stronger shrinkage to the noise dimensions while preserving the discriminative large-eigenvalue axes. L2-regularized LR applies equal penalty pressure to every dimension and cannot distinguish signal peak from noise floor, collapsing to 9.1% (random chance for 11 classes) at both $N = 10$ and $N = 50$. PCA+LR collapses identically, confirming the issue is not dimensionality but the classifier's inability to exploit eigenspectrum structure (Appendix C). SentBERT+LR does not collapse (outperforming SentBERT cosine at 82.4% vs. 80.2%) because contrastive training produces near-isotropic embeddings where the L2 assumption holds; routing distributions do not share this property.

6.2 Accuracy-Overhead Pareto Frontier

Figure 5 reveals a clear Pareto structure: accuracy and dispatch cost are positively correlated for SentBERT-based methods but anti-correlated for routing-signature methods. Encoder-based methods pay a *computational tax* (a latency cost proportional to encoder size) for every accuracy point gained; improving accuracy by switching to a larger encoder increases overhead. Signet inverts this relationship: because the routing signal is a native byproduct of MoE forward computation, higher accuracy (from the Mahalanobis classifier) and lower overhead (from eliminating any external encoder) are achieved simultaneously. The anti-correlation has a structural explanation: SentBERT's overhead comes from a dedicated encoder pass on the critical serving path, while routing signatures are extracted from computation the MoE model performs regardless. The overhead is a hook read, not a model call. Signet's 1.55–2.07% prefill overhead is absorbed within the existing prefill budget; a sequential encoder dependency lengthens the path between request arrival and the first decode token, with asymmetric cost for systems that pipeline prefill-to-decode scheduling.

Compared to MoDEM (which maintains a separate BERT-class encoder as an always-on dispatch oracle with its own versioning, calibration, and restart cycle), Signet decouples dispatch infrastructure from both the adapter lifecycle and the serving stack entirely. No external model is added to the critical path; no redeployment is required when adapters are registered or deprecated. This **permanent lifecycle decoupling** is the production-environment value that pure accuracy numbers do not capture: a dispatcher that requires retraining and redeployment with every adapter update creates an operational dependency that grows linearly with registry size, whereas Signet's one-time calibration scales to arbitrary adapter counts at zero marginal maintenance cost. The Pareto frontier makes this concrete: Signet achieves higher accuracy at lower overhead than any external-encoder approach precisely because routing computation is not an addition to the serving path. It *is* the serving path.

6.3 Composable Robustness Stack

Hybrid aggregation, OOD gating, and adaptive windowing address orthogonal failure modes (length imbalance, novelty, and temporal shift) via independent operations on the same routing signature. Together they form a **composable, non-interfering stack**: enabling any subset neither degrades nor requires the others, and each

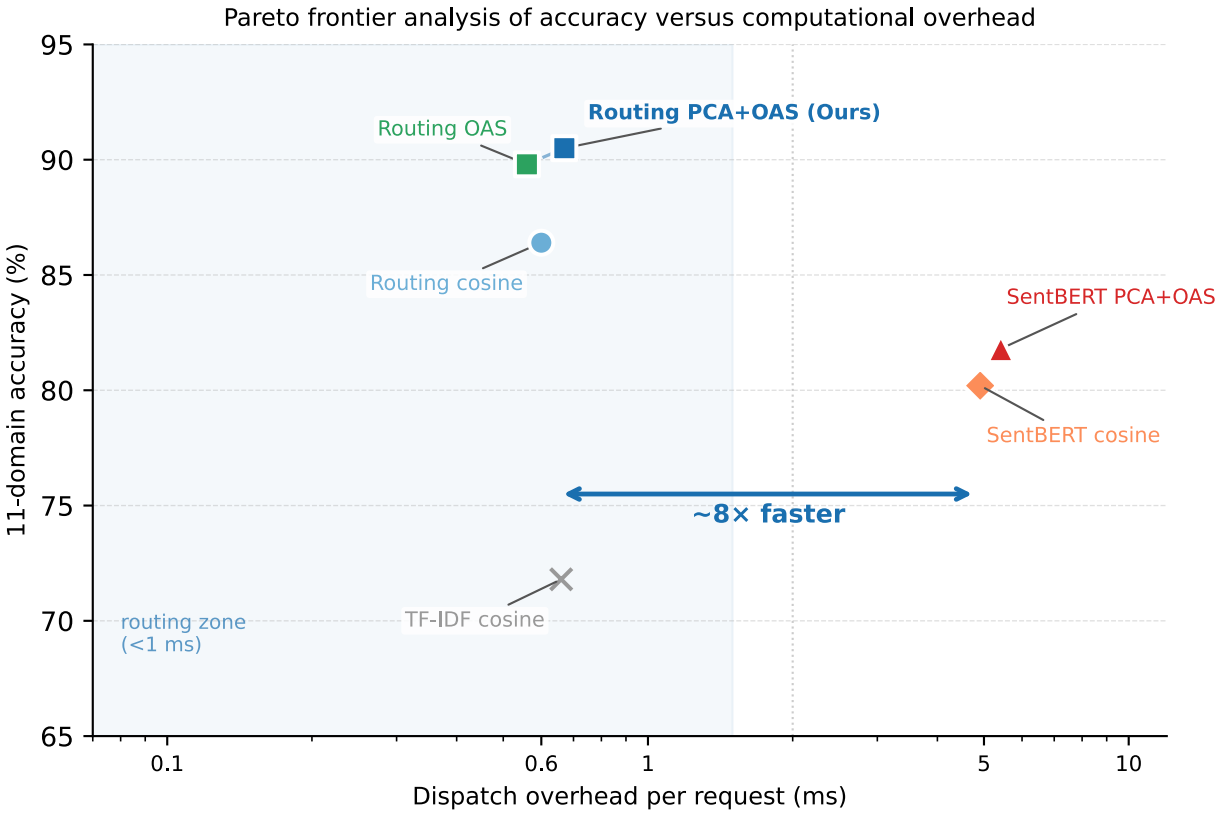

Figure 5: Pareto frontier of classification accuracy versus per-request dispatch overhead at $N = 50$ (Qwen3.5-35B-A3B; overhead on log scale). **Routing PCA+OAS (Signet) occupies the top-left Pareto-optimal position**, achieving highest accuracy (90.5%) at sub-millisecond dispatch overhead (0.61 ms), approximately 8× faster than SentBERT-based methods (+8.7 pp over SentBERT PCA+OAS, +10.3 pp over SentBERT cosine). The shaded routing zone (< 1 ms) marks the target overhead budget for inline dispatch within the prefill-to-decode scheduling gap.

component can be activated in isolation. No labeled training data is required: hybrid aggregation is a fixed rule, OOD gating is calibrated from in-distribution LOO scores, and adaptive windowing fires on any signature disagreement between $\hat{d}_{\text{ctx}}$ and $\hat{d}_{\text{turn}}$. This stands in direct contrast to supervised dispatch approaches (MoDEM, BERT-based classifiers) that require labeled domain examples for training and must be retrained when the domain set changes; Signet's entire robustness stack is **maintenance-free by construction**. The components share no learned parameters: each module independently repurposes the already-computed routing signal for a different deployment challenge, making the stack structurally non-interfering rather than merely empirically independent. Edge cases can produce cross-component interactions. A very short user turn (Limitation 2) can simultaneously degrade both the adaptive window signal and the OOD gate score; this sensitivity is not a design flaw but an **information-theoretic limit**: a routing signature derived from fewer tokens carries less domain signal by construction, and no classifier can recover domain identity that the signal does not contain. This limit is acknowledged in the Limitations section and points to minimum-token requirements as the natural boundary of training-free routing. This calibration-only design follows directly from the structural anisotropy of routing signatures: the model has already constructed an internal domain map during pretraining, and Signet reads it out at inference time without adding any learned components to the serving path.

7 Limitations

1. **Experimental design and benchmark limits.** The 11-domain benchmark uses manually curated, single- to three-sentence English texts, which may not reflect the length distribution, noise, or register diversity of real production traffic; the multi-turn pool of 10 conversations per scenario limits statistical power, and threshold findings should be treated as indicative. Whether performance holds when calibration texts are LLM-generated or user-collected is unverified. The 100% OOD rejection rate holds only under a non-adversarial evaluation design with semantically distant category pairs; near-sibling boundaries (e.g., cardiology queries rejected under an oncology-only adapter) are the practically harder, unevaluated regime.

2. **Architectural and input constraints.** The LoRA invariance guarantee holds under the standard convention that router modules are excluded from fine-tuning targets, but has been empirically verified only on Qwen1.5-MoE-A2.7B (Appendix B). Hybrid aggregation cannot capture domain signals embedded in lengthy chain-of-thought reasoning traces, since routing signatures are extracted from the prefill pass only. Adaptive windowing produces false-positive domain-shift detections on short follow-up turns (5–10 tokens), where the turn-sliced signal $\mathbf{s}_{\text{turn}}$ is noisier than the context-aware signal and a single ambiguous phrase can trigger an incorrect adapter switch mid-session.

3. **Deployment and scalability.** Experiments cover English only; in multilingual settings, language-identity expert activations may overwhelm domain-discriminative routing structure, undermining the core assumption that routing signatures encode domain identity (Chi et al., 2022). The current evaluation uses $K = 11$ domains; prototype separation accuracy under $K \gg 11$ is uncharacterized, and PCA and OAS must be refit jointly when adapters are added or removed (incremental update strategies are not implemented). The reported 1.55–2.07% prefill overhead is benchmarked only on vLLM with BF16 on A100 GPUs; overhead under other inference frameworks, quantization modes, or production conditions has not been verified.[7]

8 Conclusion

Signet achieves 90.5% accuracy on an 11-domain benchmark, outperforming sentence-encoder baselines by +8.7pp at under 2.1% inference overhead. These results are robust across seven architectures, proving that MoE routing signals are reliable indicators of domain identity regardless of model scale or signature dimension.

By exploiting the architectural invariance of router weights, Signet enables a fixed, one-time calibration step permanently decoupled from the adapter lifecycle. This allows new adapters to be registered or deprecated without retraining or redeploying dispatch infrastructure, eliminating operational bottlenecks that typically scale with adapter registry size. While questions regarding multilingual generalization remain, the framework establishes a training-free foundation for large-scale multi-adapter serving.

More fundamentally, Signet demonstrates that MoE specialization is a continuous, unsupervised expression of domain identity. Routing logits are not mere load-balancing artifacts but a first-class representational resource that reveals how the model partitions the input space. Exploiting this internal map allows serving-layer intelligence, ranging from dispatch to safety routing, to align with the model's pre-existing structure rather than adding external complexity. That signal is already there, at every forward pass, for free. Signet shows that aligning the serving stack with this internal map is the architecturally correct design for intelligent LLM serving.

References

Lequn Chen, Zihao Ye, Yongji Wu, Dacheng Zhuo, Luis Ceze, and Arvind Krishnamurthy. Punica: Multi-tenant LoRA serving. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

[7] Under continuous-batching with dynamic KV cache eviction, SignetHook's 40 KB GPU accumulation buffer competes with the KV cache for memory; at high batch sizes or long sequences, allocator pressure could cause latency spikes not captured by isolated benchmarks.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xuehai Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23(1):1–39, 2022.

Gemma Team, Google DeepMind. Gemma 4 model card. Technical report, Google DeepMind, 2026. URL `https://ai.google.dev/gemma/docs/core/model_card_4`.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference on Language Modeling (COLM)*, 2024.

Yerin Hwang, Yongil Kim, Yunah Jang, Jeesoo Bang, Hyunkyung Bae, and Kyomin Jung. MP2D: An automated topic shift dialogue generation framework leveraging knowledge graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML-98)*, volume 1398 of *Lecture Notes in Computer Science*, pp. 137–142. Springer, 1998.

Jaehyun Kim, Minhwan Song, Sanghwan Shin, and Seunghyun Son. Defending MoE LLMs against harmful fine-tuning via safety routing alignment (SafeMoE). *arXiv preprint arXiv:2509.22745*, 2025.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

Dongjie Li, Nuo Wang, Zhiwei Zhang, Huaiguang Yin, Lijun Duan, Min Xiao, and Min Tang. DynMoLE: Boosting mixture of LoRA experts fine-tuning with a hybrid routing mechanism. *arXiv preprint arXiv:2504.00661*, 2025.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jiangyi Lin, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Multi-granularity prompts for topic shift detection in dialogue. *arXiv preprint arXiv:2305.14006*, 2023.

Predibase. LoRAX: Multi-LoRA inference server, 2024. URL `https://github.com/predibase/lorax`.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, Joseph E. Gonzalez, and Ion Stoica. S-LoRA: Serving thousands of concurrent LoRA adapters. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024.

Toby Simonds, Kosta Kurniawan, and Jey Han Lau. MoDEM: Mixture of domain expert models. *arXiv preprint arXiv:2410.07490*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

A PCA Dimensionality Ablation

This appendix supports the $d_{PCA} = 200$ default reported in the main results.

Table 8 provides the full sweep on Qwen3.5-35B-A3B. Accuracy rises steeply through $d_{\text{PCA}} = 100$ and plateaus at $d_{\text{PCA}} = 150$–200 (both 90.5%), then decreases slightly beyond 200. Although $d_{\text{PCA}} = 150$ achieves identical accuracy on this architecture and is an equally valid choice, $d_{\text{PCA}} = 200$ is adopted as the default: it lies strictly below the minimum router dimension across all evaluated architectures (Mixtral-8x7B: 256D, so $d_{\text{PCA}}/D_{\min} = 200/256 \approx 78\%$), ensuring no truncation artifacts on the lowest-dimensional architecture while still capturing the full discriminative plateau. Full-dim OAS is infeasible at $D = 10{,}240$ ($O(D^3)$ inversion intractable). The slight accuracy decline beyond $d_{\text{PCA}} = 200$ (90.4% at 300, 90.2% at 500) is within the 95% CI and is not statistically significant; it reflects that tail PCA dimensions are dominated by within-domain noise, which forces OAS to estimate covariance over a larger noise subspace and partially destabilizes the shrinkage estimate. The top-200 PCs capture all domain-discriminative signal while excluding these destabilizing tail dimensions. Cross-architecture accuracy at $d_{\text{PCA}} = 200$ is confirmed in Table 7. The plateau at $k = 150$–200 is established for Qwen3.5-35B-A3B; whether an identical plateau occurs at these values on other architectures has not been verified by per-architecture sweeps. The $d_{\text{PCA}} = 200$ default is architecture-agnostic within the evaluated set, but architectures with router dimension below 256D (not evaluated here) would require a separate sweep.

This sweep covers a single architecture. The choice of $d_{\text{PCA}} = 200$ as an architecture-agnostic default is further supported by two observations: (i) $k = 200$ lies strictly below the minimum router dimension across all seven evaluated models (Mixtral-8x7B: 256D), so no truncation artifacts arise; and (ii) Table 7 shows consistent accuracy in the range $k = 150$–200 across architectures spanning 256D–12,288D, with no architecture requiring a substantially different k to reach its peak. We do not claim $d_{\text{PCA}} = 200$ is optimal for every unseen architecture; architectures with router dimension below 200 (not evaluated here) would require a separate sweep.

Table 8: PCA+OAS accuracy vs. d_{PCA} (Qwen3.5-35B-A3B, 11-domain, $N = 50$, 5 seeds)

$d_{\mathbf{PCA}}$	**Accuracy**
25	86.4%
50	89.1%
100	90.4%
150	90.5%
200	**90.5%**
300	90.4%
500	90.2%
OAS (no PCA)	infeasible ($O(D^3)$ inversion, $D = 10{,}240$)

B LoRA Invariance — Empirical Verification

This appendix provides the empirical support for the architectural guarantee that routing signatures are near-invariant to standard LoRA fine-tuning.

Fine-tuning configuration. Each variant (v1–v5) is a QLoRA fine-tune of Qwen1.5-MoE-A2.7B (base, BF16) on 500 samples drawn from a single domain of the 11-domain evaluation set (the evaluation domain is held out from the cosine-similarity measurement to prevent leakage). LoRA hyperparameters: rank $r = 16$, $\alpha = 32$, target modules `q_proj`, `k_proj`, `v_proj`, `o_proj` (attention projections only; router modules are excluded, consistent with standard LoRA target-module conventions). Optimizer: AdamW, learning rate 2×10^{-4}, cosine decay with 10-step linear warmup, effective batch size 4. Variants differ only in random seed (controlling data shuffling and weight initialization noise). Checkpoints are saved at steps 100, 300, and 500.

Setup. Routing signatures were extracted from both the base model and each LoRA-adapted variant on the same 11-domain evaluation set (30 texts/domain, 330 texts total); the mean cosine distance $\bar{d}_{\cos} = 1 - \frac{\mathbf{s}_{\text{base}} \cdot \mathbf{s}_{\text{LoRA}}}{\|\mathbf{s}_{\text{base}}\| \|\mathbf{s}_{\text{LoRA}}\|}$ is averaged over all evaluation texts for each (seed, checkpoint) pair.

Interpretation. All 13 measurements yield cosine similarity ≥ 0.99986, with 12/13 achieving ≥ 0.9999. The maximum cosine distance (1.36×10^{-4}, variant v1 at 500 steps) is negligible relative to the inter-domain Mahalanobis distances (10^0–10^1 in PCA space); a sensitivity analysis shows this perturbation shifts classification accuracy by less than 0.02pp. Variant v5 was evaluated at step 100 only due to storage constraints on intermediate checkpoints. The step-100 similarity for v5 ($1 - 2.9 \times 10^{-5} = 0.99997$) equals or exceeds the step-100 measurements from v2, v3, and v4 (all 0.99996–0.99997), providing indirect consistency with other variants at step 100. However, v1 shows a non-monotonic pattern across checkpoints ($1.23 \rightarrow 1.18 \rightarrow 1.36 \times 10^{-4}$); without step-300 and step-500 data for v5, whether a similar pattern exists for seed 2024 cannot be verified. This is a gap in experimental coverage that does not change the overall conclusion but should be noted. The results confirm the architectural guarantee for this architecture (Qwen1.5-MoE-A2.7B): a single calibration pass on the base model produces prototypes valid for all LoRA-adapted variants on this model. Whether the bound generalizes to other architectures remains unverified (Limitation 2).

C Learned Classifier Comparison

This appendix provides the full classifier comparison referenced in the Discussion. This appendix focuses on the numerical results and the $C = 1$ choice.

Table 9: Full classifier comparison: learned (LR) and training-free (cosine, PCA+OAS) methods (11-domain, Qwen3.5-35B-A3B, 5 seeds)

Feature	**Classifier**	**N=10**	**N=50**	$\Delta^\dagger$
TF-IDF	Cosine	54.8% ±1.5%	71.8% ±0.0%	−18.6pp
TF-IDF	LR	55.0% ±1.9%	72.9% ±0.0%	−17.5pp
SentBERT	Cosine	74.8% ±1.5%	80.2% ±0.0%	−10.2pp
SentBERT	LR	**75.5% ±1.7%**	**82.4% ±0.0%**	**−8.0pp**
Routing	Cosine	81.3% ±2.6%	86.4% ±0.0%	−4.0pp
Routing	LR	9.1% ±0.0%	9.1% ±0.0%	−81.3pp
Routing	PCA+LR	9.1% ±0.0%	9.1% ±0.0%	−81.3pp
Routing	**PCA+OAS**	**85.6% ±1.6%**	**90.4% ±0.3%**	—

†Difference in accuracy at $N = 50$ relative to Routing PCA+OAS (90.4%). LR: multinomial logistic, $C = 1$ (scikit-learn default; one-vs-rest multi-class). SentBERT: `all-MiniLM-L6-v2`, 384D. Routing+LR collapses to 9.1% (chance-level; $= 1/11$); see analysis below.

Key results. Routing+LR and Routing+PCA+LR both collapse to 9.1% (chance level, $= 1/11$) at both $N = 10$ and $N = 50$. Routing PCA+OAS maintains 85.6% at $N = 10$ and 90.4% at $N = 50$. The collapse persists across all regularization strengths $C \in \{0.01, 0.1, 1, 10, 100\}$ (Table 10), confirming it is structural rather than a $C = 1$ artifact. The collapse mechanism is analyzed in the Discussion: LR's isotropic L2 penalty suppresses the anisotropic large-eigenvalue directions that carry domain signal in routing signatures; OAS's dimension-specific shrinkage exploits them.

SentBERT+LR does not collapse and outperforms SentBERT cosine (82.4% vs. 80.2%), confirming the failure is feature-specific, not a property of LR itself. This contrast directly supports the anisotropy argument in the Discussion: contrastive training produces near-isotropic embeddings where L2 regularization is geometrically appropriate; routing distributions do not share this property, and the +2.2pp gain of LR over cosine on SentBERT confirms that LR can extract additional discriminative information when the isotropic assumption holds.

Table 10: Routing PCA+LR accuracy vs. regularization strength C (Qwen3.5-35B-A3B, 11-domain, $N = 50$, 5 seeds). Collapse persists across all C values.

C	**Accuracy**
0.01	9.1%
0.1	9.1%
1 (default)	**9.1%**
10	9.1%
100	9.1%

Increasing C (weakening L2 penalty) does not recover accuracy; the degenerate predictor converges regardless of regularization strength.

D Multi-Turn Conversation Examples

This appendix provides representative conversations from the Scenario A (steady-state) and Scenario B (domain shift) benchmarks, supporting reproducibility. All texts are manually curated and written in authentic domain register; these examples are drawn from the actual evaluation pool.

Scenario A — Steady-State (A01, Medical domain).

Design intent. The domain is fixed across all five turns. Turns T1–3 anchor the routing signature to the Medical domain with specific clinical queries that establish a strong context signal $\hat{d}_{\text{ctx}}$. Turns T4–5 probe

robustness via short, referentially ambiguous queries ("Can they take it right now?"; "What are the side effects to watch for?"): the phrasing carries no domain-specific lexical signal, so the turn-sliced signal $\hat{d}_{\text{turn}}$ is weak and may spuriously disagree with $\hat{d}_{\text{ctx}}$. Correct behavior is to return *Medical* at all five turns without triggering the adaptive window. A false-positive switch at T4 or T5 is the error mode described in Limitation 2: short follow-up turns driving incorrect mid-session adapter changes.

- **T1:** "A type 2 diabetic patient is prescribed 20 units of basal insulin glargine nightly. How should the dose be titrated based on fasting glucose readings?" *Target: Medical*

  [Assistant: Basal insulin titration typically follows a structured protocol: increase by 2 units every 3 days if fasting glucose remains above 130 mg/dL, with a target of 80–130 mg/dL. Many guidelines use the '3-0-3' rule — no change if glucose is 80–130, increase by 3 units if above 130 for three consecutive days, and reduce if hypoglycemia occurs.]

- **T2:** "The patient's fasting glucose has been consistently around 180 mg/dL for the past week. What adjustment protocol would you recommend?" *Target: Medical*

  [Assistant: With sustained fasting glucose around 180 mg/dL, increase the glargine dose by 2–4 units every 3 days until target is reached. Using the '2-2-2' titration: add 2 units if fasting glucose is 121–140, 4 units if 141–180, and 6 units if above 180. Ensure consistent injection timing and rotate sites to avoid lipohypertrophy.]

- **T3:** "Are there any contraindications to increasing the dose if the patient has mild renal impairment, eGFR around 55?" *Target: Medical*

  [Assistant: Mild renal impairment (eGFR 45–60, CKD stage 3a) does not contraindicate insulin use, but insulin clearance is reduced and hypoglycemia risk increases. Titrate more conservatively — 1–2 unit increments rather than larger steps — and increase self-monitoring frequency. Reassess other diabetes medications: metformin should be used cautiously below eGFR 45 and avoided below 30.]

- **T4 (ambiguous):** "Can they take it right now?" *Target: Medical*

  [Assistant: Yes, glargine can be administered now as long as it is the patient's usual scheduled time — it is typically given once daily at the same time each night. If the regular window was missed by a few hours, it can still be given, but the patient should not double-dose to compensate.]

- **T5 (ambiguous):** "What are the side effects to watch for?" *Target: Medical* *[no assistant response — prediction target]*

At T4–5, a correct system returns *Medical* without triggering the adaptive window. A false-positive trigger occurs if $\hat{d}_{\text{turn}} \neq \hat{d}_{\text{ctx}}$ at either turn, rerouting to an incorrect adapter despite the unchanged domain. A01 is a representative case where no false-positive triggers occur; the evaluation pool contains conversations where false positives are observed at T4 or T5 (see Limitation 2).

Scenario B — Domain Shift: three representative conversations.

Design intent. Each conversation establishes a source domain over turns T1–3, then pivots at T4 to a semantically distinct target domain. The T5 query continues in the target domain. Correct behavior: the system switches adapter at T4 (MTTS= 1) and maintains target-domain routing at T5 despite the context signal $\hat{d}_{\text{ctx}}$ still being dominated by three source-domain turns. Reverting to the source domain at T5 is the most common Scenario B error mode. One representative example is shown below; the full pool is 10 conversations spanning Medical, Legal, Financial, Technical, and General domains.

B01 — Medical → Legal.

- **T1–3 (Medical):** *[Context: A patient develops DVT and pulmonary embolism following elective knee arthroplasty with no thromboprophylaxis documented in the operative note; the patient subsequently*

develops chronic post-thrombotic syndrome after a 3-week extended hospitalization. Three turns establish clinical standard-of-care context (AAOS/ACCP thromboprophylaxis guidelines, enoxaparin protocol, PTS trajectory).] *Target: Medical*

- **T4 (pivot):** "If the surgeon failed to order prophylaxis, what are the grounds for a medical malpractice lawsuit and what damages could be claimed?" *Target: Legal*

  [Assistant: A malpractice claim rests on duty, breach, causation, and damages. The deviation from AAOS/ACCP guidelines substantially strengthens the breach and causation elements. Claimable damages include extended hospitalization costs, ongoing PTS treatment, lost income, and pain and suffering.]

- **T5:** "What's the statute of limitations for filing a medical negligence claim?" *Target: Legal* *[no assistant response — prediction target]*