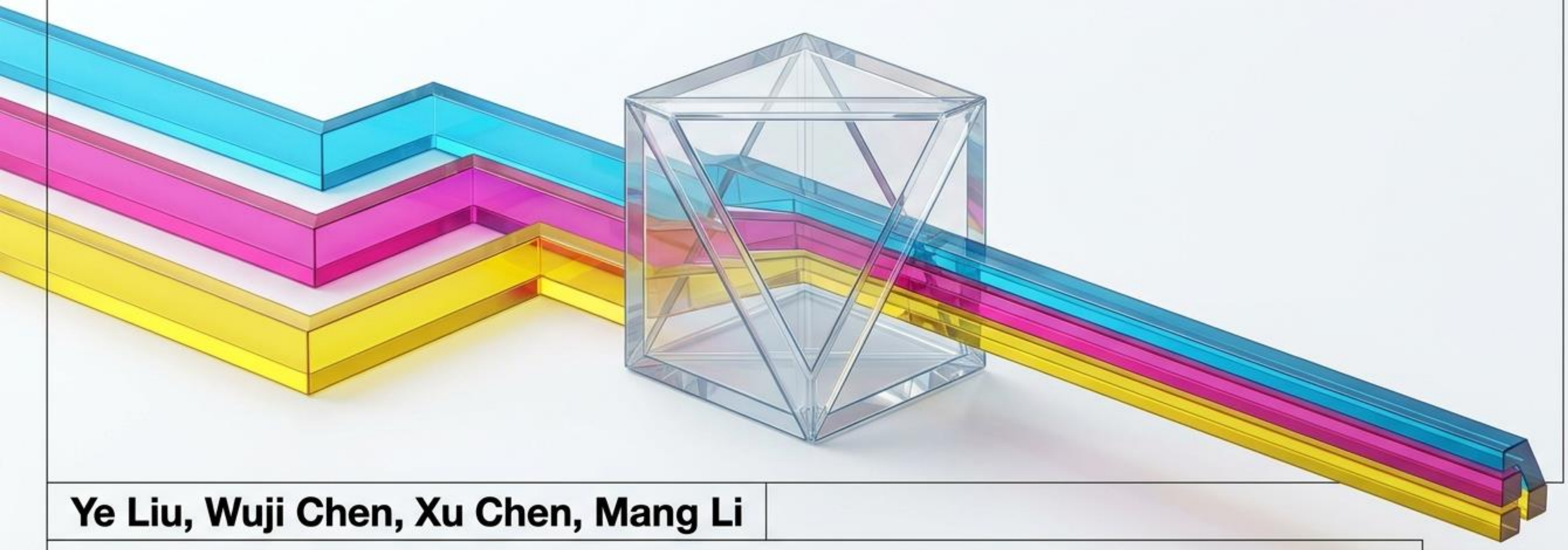
Orchestrating Heterogeneous Experts

A Scalable MoE Framework with Anisotropy-Preserving Fusion



Ye Liu, Wuji Chen, Xu Chen, Mang Li

Institute of Intelligent Technology, Alibaba International Digital Commerce Group

WWW '26 (Dubai, UAE)

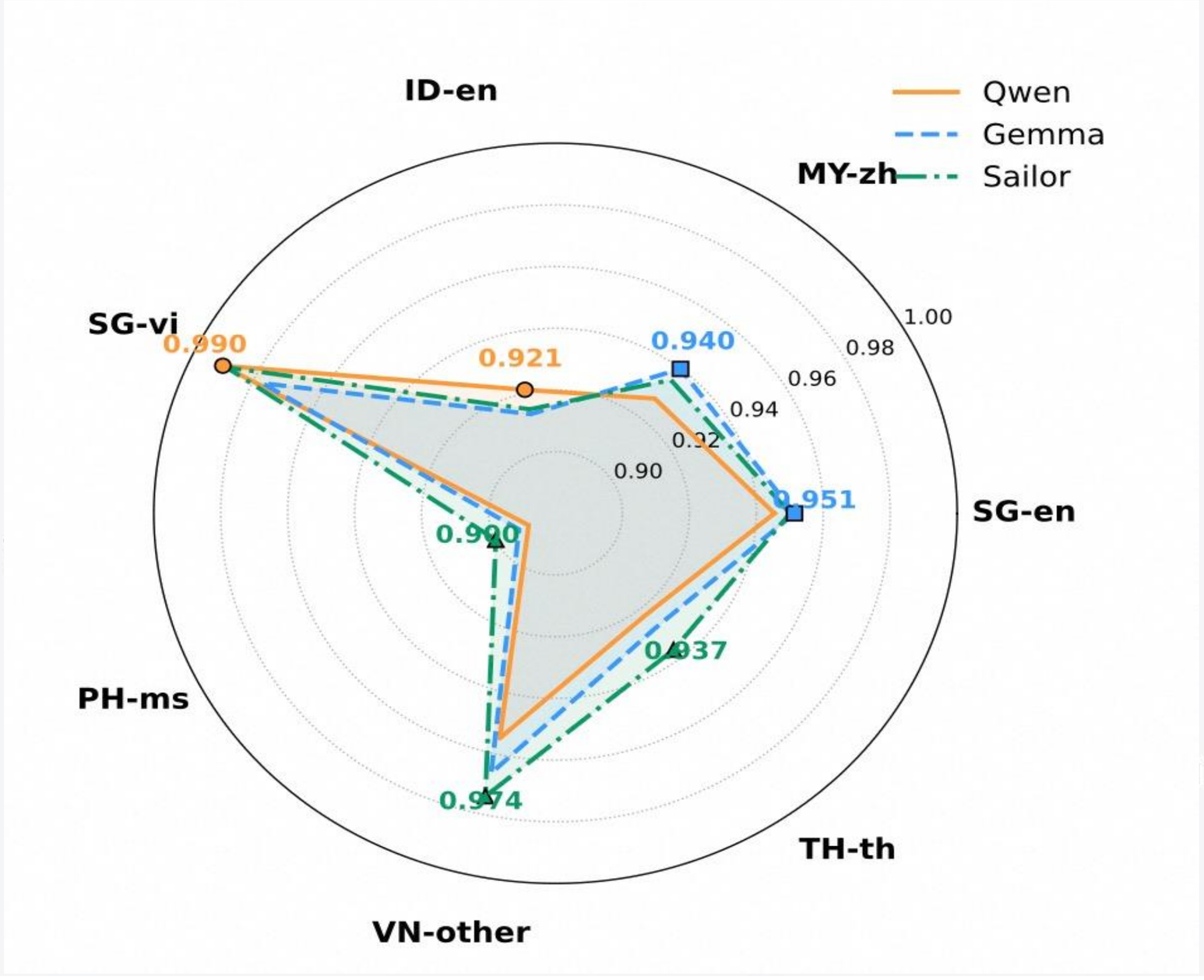
The Challenge

Extreme linguistic diversity meets intense query ambiguity.

The Flaw of the Status Quo

Scaling monolithic Pre-trained Language Models (PLMs) leads to **Capacity Dilution**. Forcing a single model to master complex English reasoning alongside highly localized Asian syntax results in destructive performance trade-offs.

Identical Fine-Tuning Yields Distinct Regional Mastery

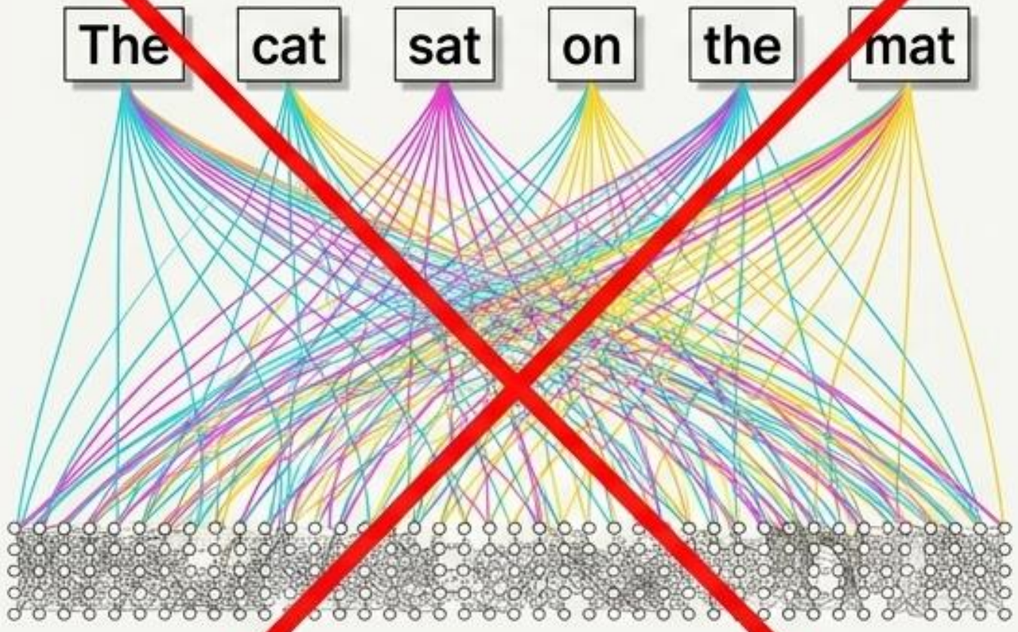


Differences in pre-training corpora and **vocabulary design** cause models to internalize **distinct, complementary regional capabilities.**

Why force one monolithic model to learn everything when we can orchestrate complementary experts?

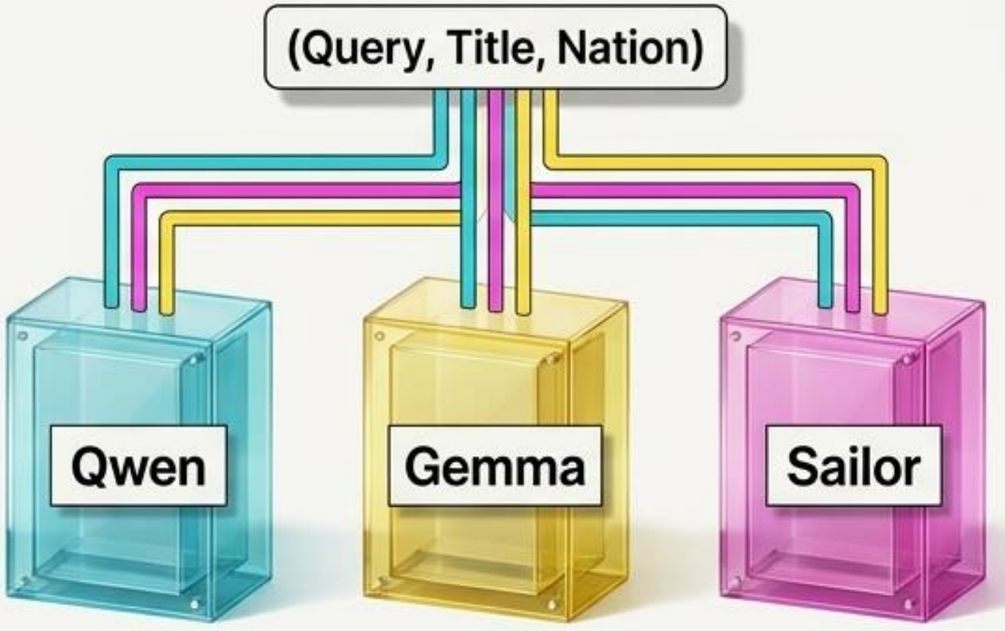
Shifting to Request-Level Expert Orchestration

Token-Level MoE



Scaling One Model = High latency, steep curve of diminishing returns. Computationally heavy, requires continuous pre-training.

Request-Level MoE

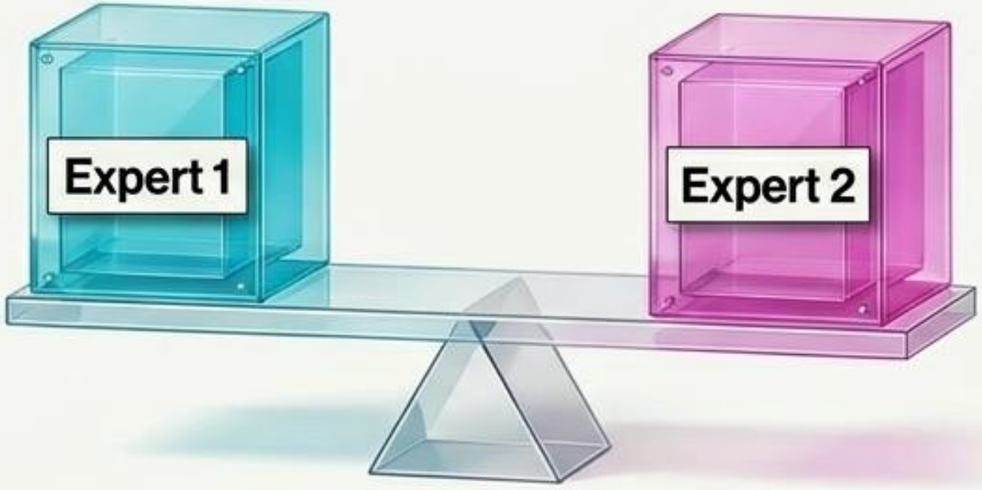


Orchestrating Many = Dynamically routes whole queries to specialized experts. Exploits diverse open-source LLMs without expensive continuous pre-training.

End-to-End Hard Routing Prevents Expert Collapse

Strategy	Flexibility	Training-Inference Match
Rule-based	Rigid	N/A
Soft Routing	High	Mismatched (requires threshold tuning) tuning)
End-to-End Hard Routing	**High**	**Perfect Match**

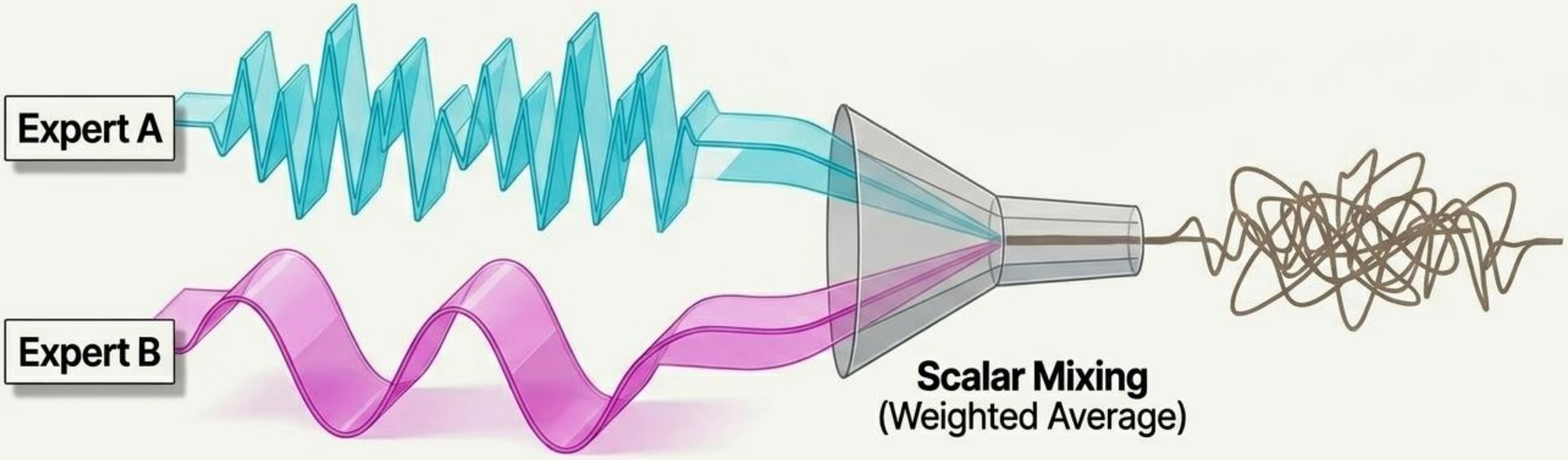
Directly selects Top-K experts (K=2) per query.



Load Balancing Loss: $\mathcal{L}_{LB} = N \cdot \sum p_i \cdot p_i$

Prevents the network from collapsing to a single easiest expert, ensuring diverse utilization.

The Fatal Flaw of Scalar Mixing: Destructive Interference



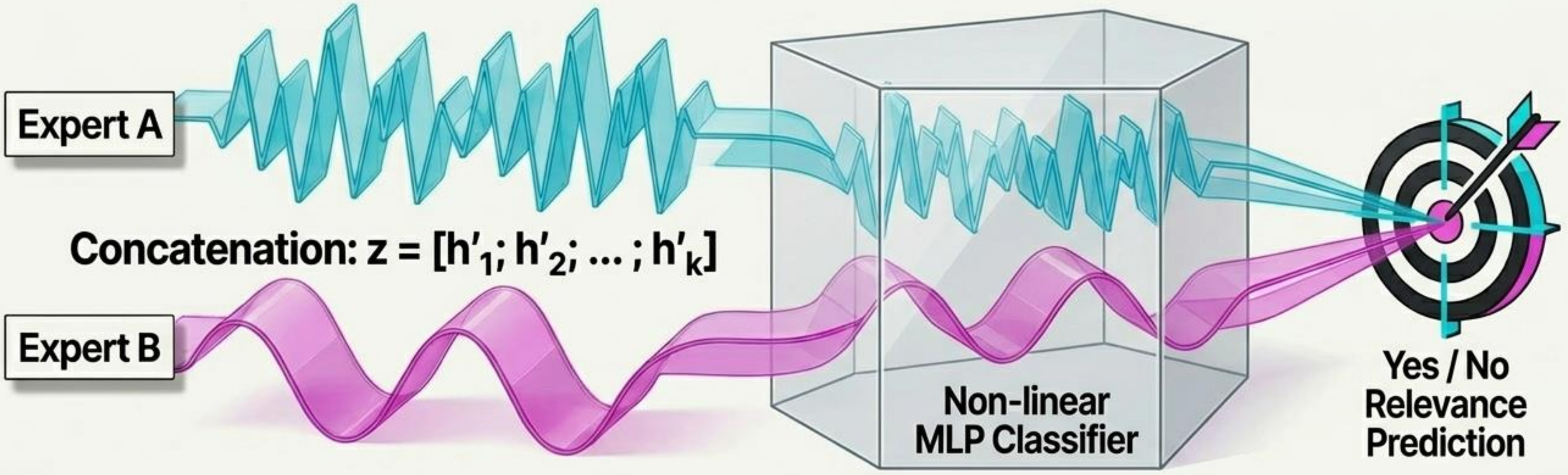
The Misalignment Problem

Heterogeneous LLMs utilize distinct tokenizers and architectures. Their latent manifolds are topologically distinct. The basis vectors are not naturally aligned.

The Result

Linearly combining embeddings collapses fine-grained semantic signals into featureless noise, destroying the exact information we aimed to capture.

Information-Preserving Concatenation

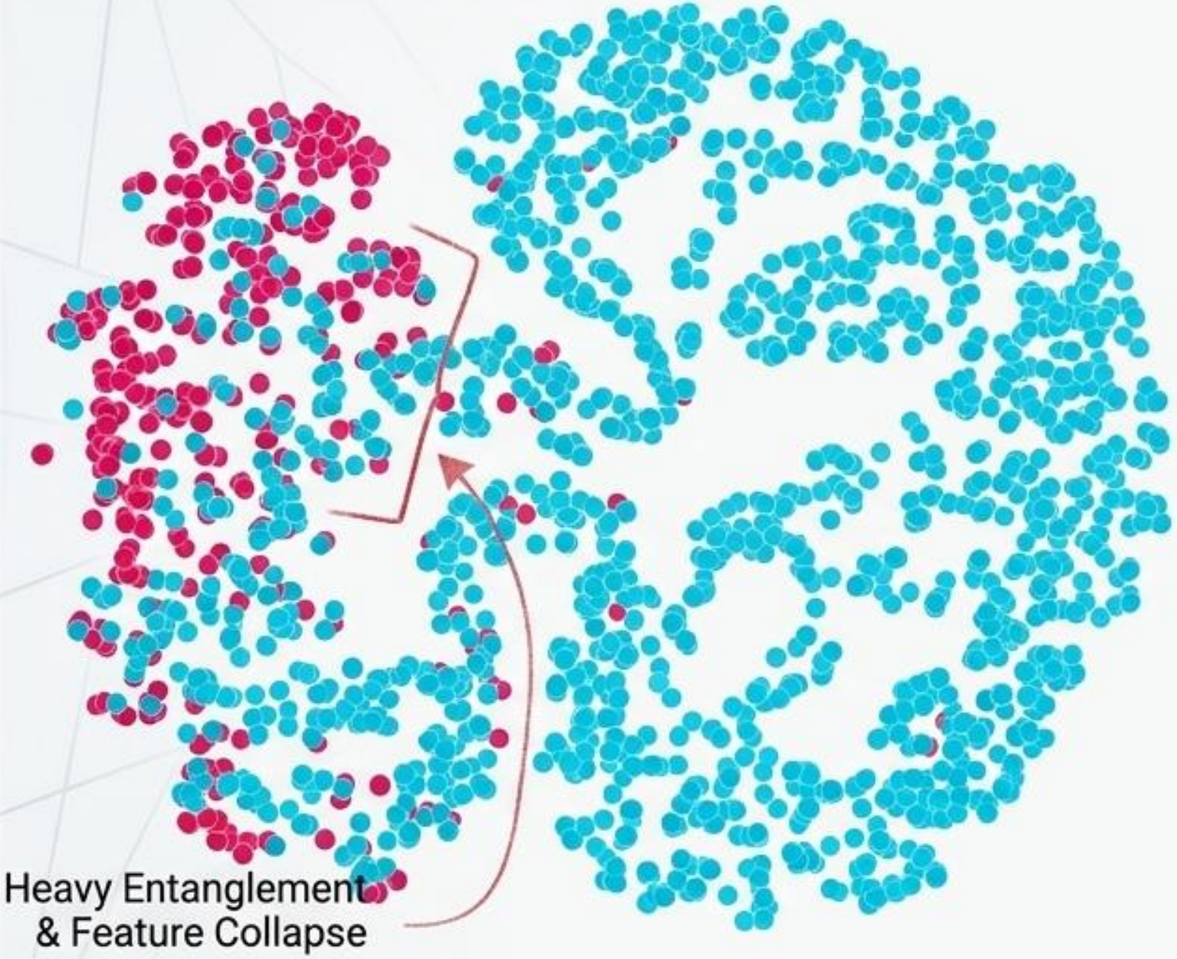


Core Innovation Construct a joint representation space by **concatenating projected embeddings**, **preserving the intrinsic manifold structure and geometry of each expert.**

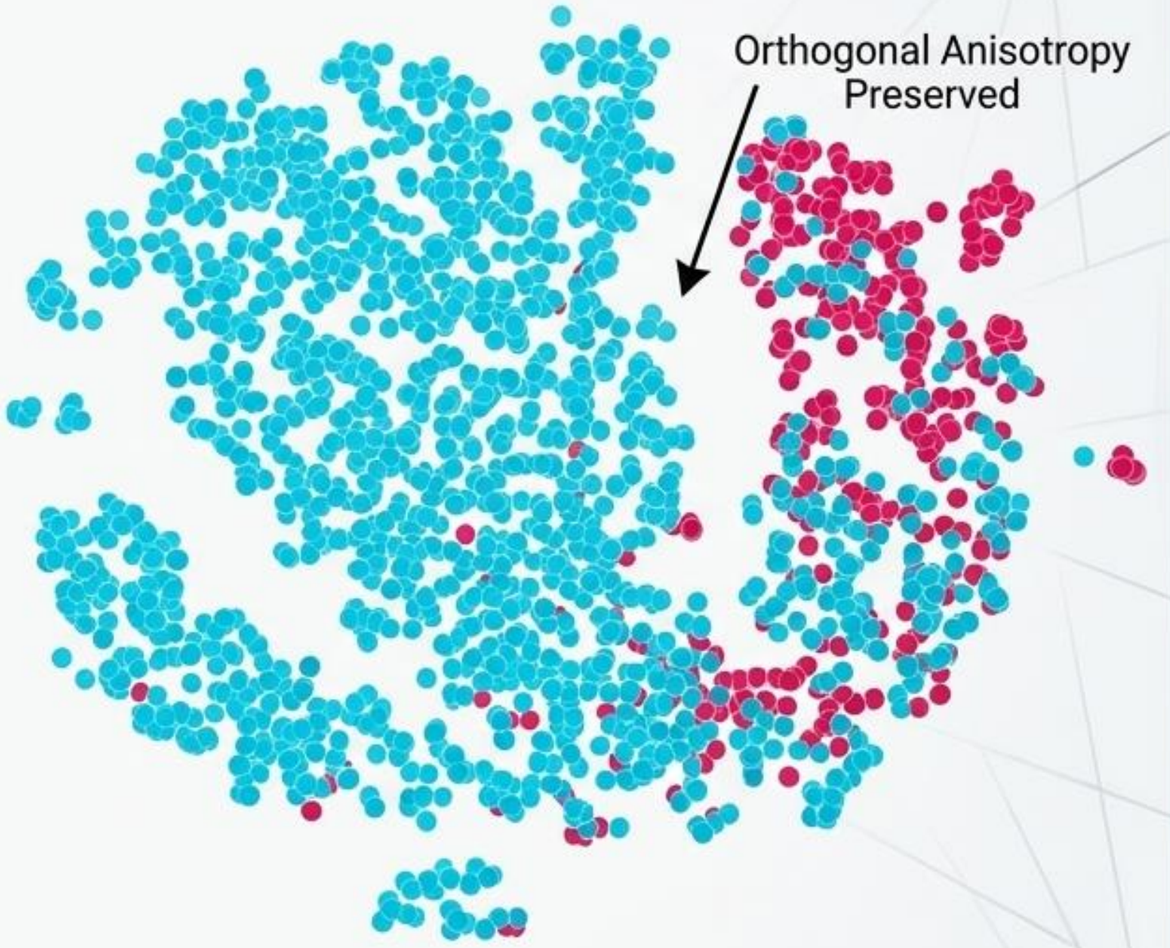
Non-linear Discrimination An MLP learns which expert to trust for **specific feature subspaces**, utilizing the sharpest features from either model **without information loss.**

Geometric Proof: Unfolding the Latent Manifold

Weighted Average: Entangled Manifold



Concatenation: Topological Detachment

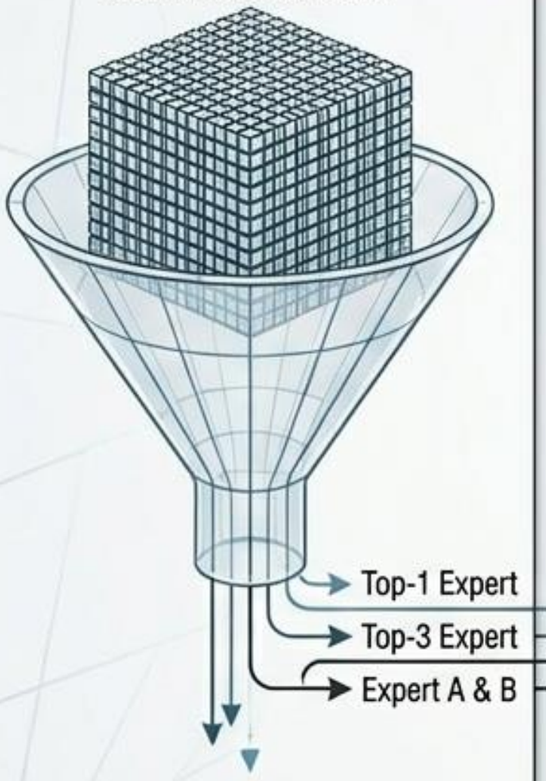


Conclusion: By preserving high-variance spikes across dimensions, the downstream classifier draws a clean, linear decision boundary, resolving the feature collapse of traditional fusion

Asynchronous Multi-Stream Batch Inference

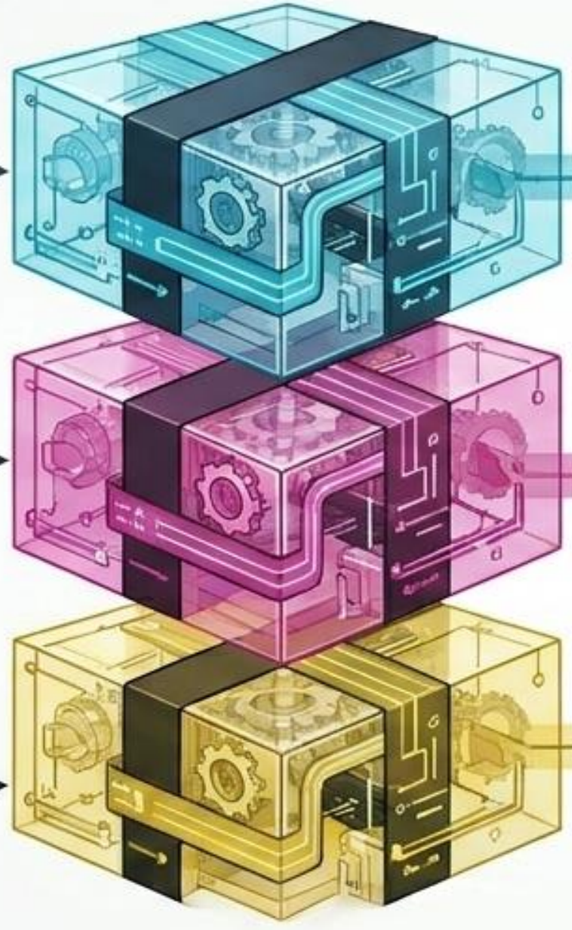
Bulk Routing

Query Logs
50M+ Records/Batch



Decoupling Compute: Routing is isolated from heavy computation, outputting sparse Top-K decisions.

Async Expert Inference



Expert A (Cyan)
90%
Executing... 15ms/batch

Expert B (Magenta)
50%
Executing... 25ms/batch

Expert C (Yellow)
25%
Executing... 40ms/batch

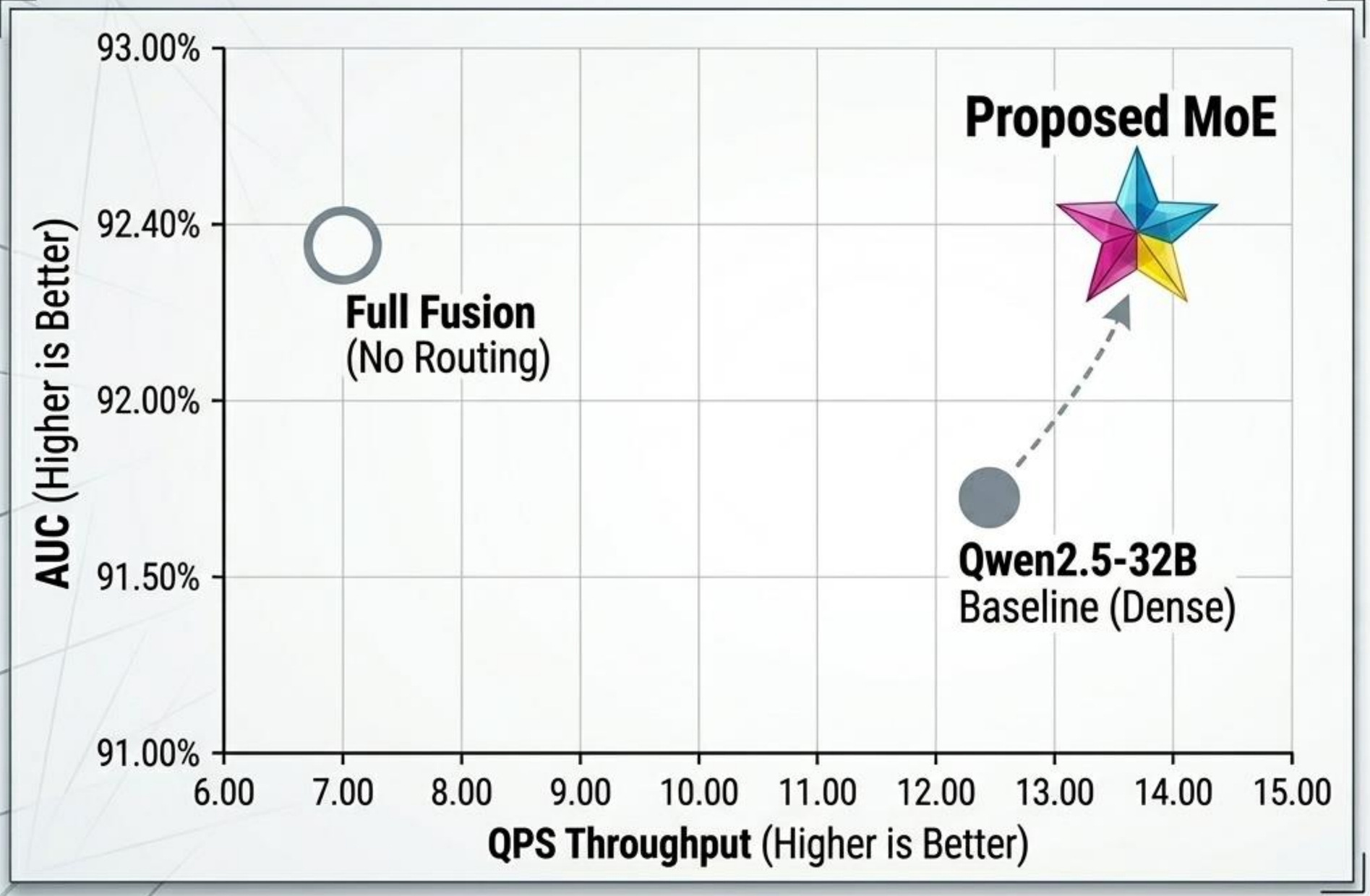
Resource-Efficient Scheduling: Experts execute independently in parallel GPU streams. Faster experts assist slower ones, maximizing cluster-level utilization.

Late-Stage Fusion

High-Throughput Result

High throughput is maintained without sacrificing relevance accuracy.

Achieving Pareto-Optimal Relevance and Throughput



+0.72%

AUC improvement over the parameter-equivalent 32B dense baseline.

9% Faster

Throughput (13.72 QPS) compared to the single large model.

Insight: "Full Fusion" underperforms Sparse Routing, proving that selecting the right experts reduces noise and improves accuracy.

Distillation Multiplier: Live Traffic Validation

Knowledge Transfer

Offline Teacher: Massive MoE Framework

Billion-Scale Relevance Distillation

Deployed Student: Compact ColBERT Model

The Deployment Strategy:
The MoE serves as an offline teacher, generating high-quality relevance labels to distill a low-latency model for live online retrieval.

Key Metric Panel 1

-5% Relative Reduction

In Bad Ratio (irrelevant results) vs. strongest single-model teacher.

Key Metric Panel 2

91.28%

Online AUC

Proving **fine-grained semantic nuances** are successfully transferred to live traffic.

Redefining Scalability in Relevance Modeling

01

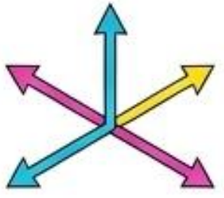
Shift the Paradigm

Orchestrate, don't just scale. Inherent regional complementary strengths bypass the need for expensive continuous pre-training.

02

Preserve the Manifold

Concatenation is mathematically superior to averaging for heterogeneous LLMs, preventing destructive interference and preserving orthogonal anisotropy.



03

Industrial Scale

A highly viable, Pareto-optimal architecture that transfers state-of-the-art multilingual understanding seamlessly into live production traffic.