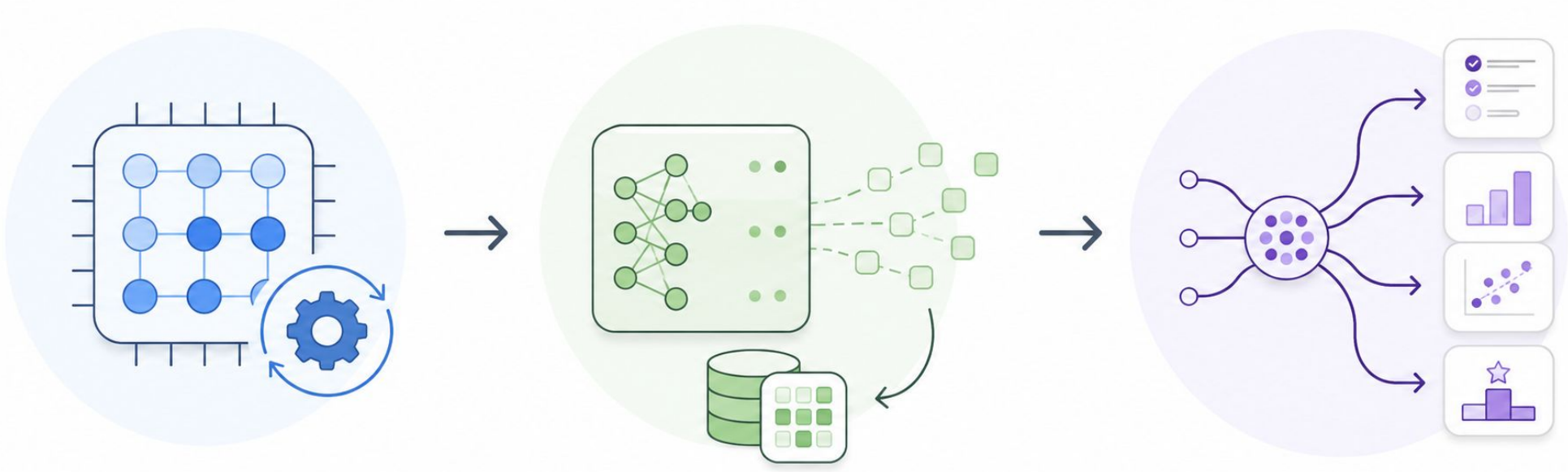
Efficiency Across the Foundation Model Lifecycle: Training, Inference, and Evaluation

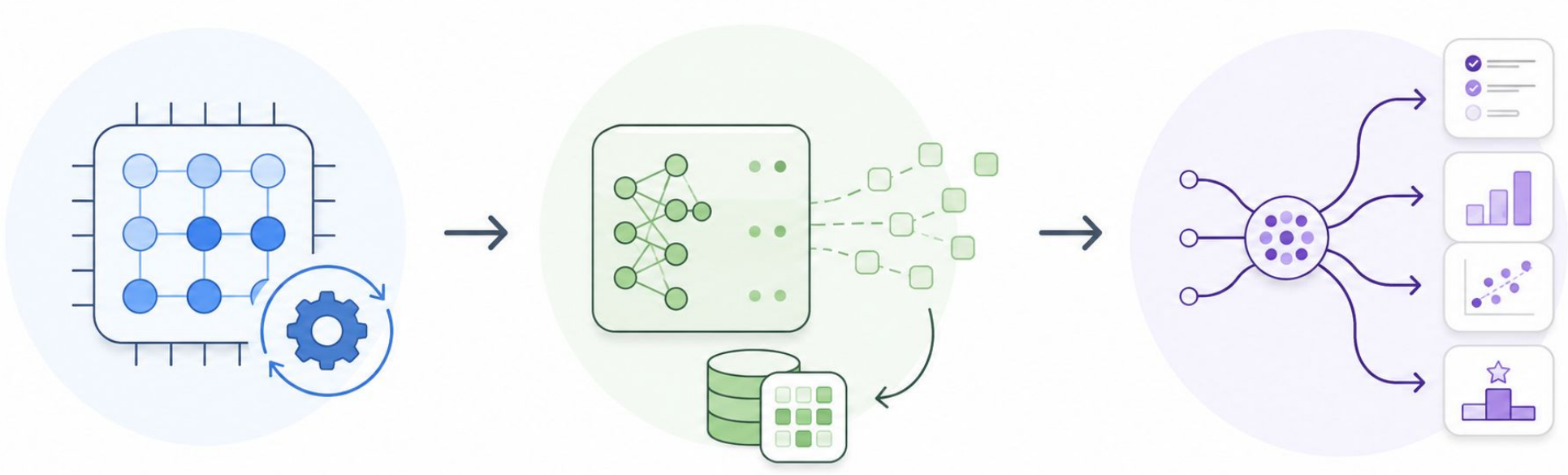
Fangzhou Wu

Efficiency Across the Foundation-Model Lifecycle



Training **Inference** **Evaluation**

Efficiency Across the Foundation-Model Lifecycle



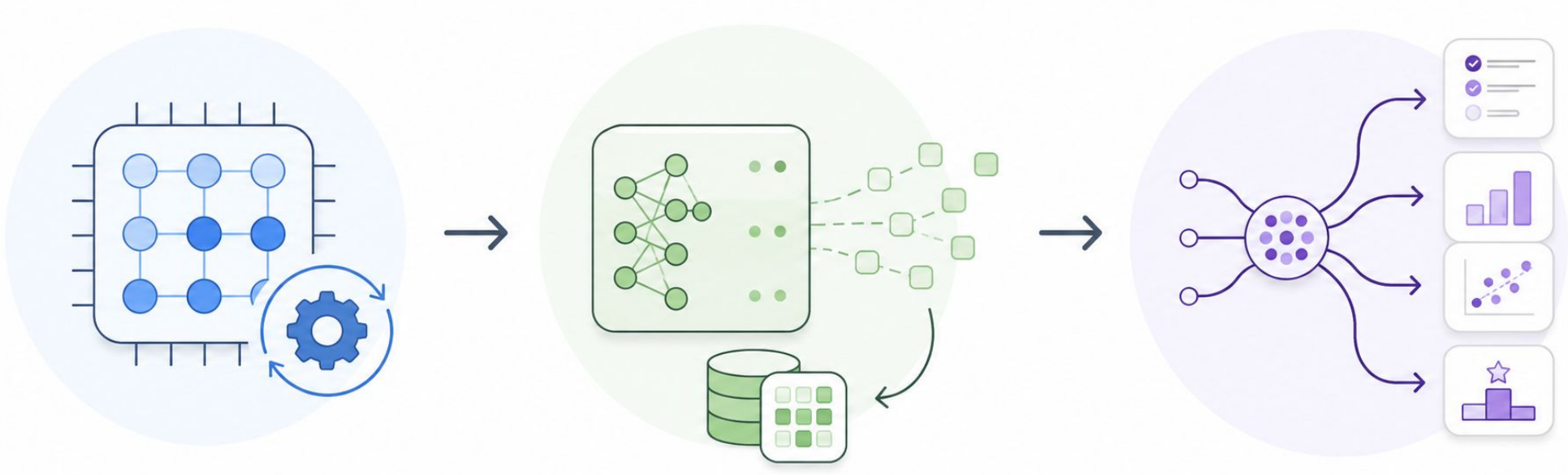
Training

How can we *train* foundation models in fewer optimization steps?

Inference

Evaluation

Efficiency Across the Foundation-Model Lifecycle



Training

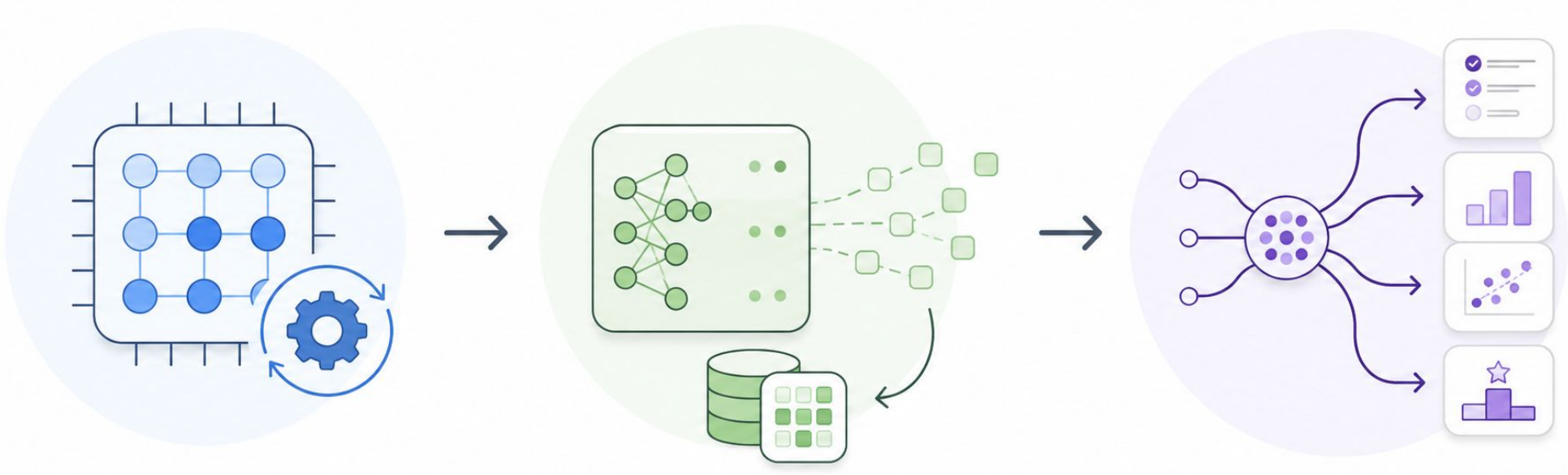
How can we *train* foundation models in fewer optimization steps?

Inference

How can we *serve* foundation models faster under memory constraints?

Evaluation

Efficiency Across the Foundation-Model Lifecycle



Training

How can we *train* foundation models in fewer optimization steps?

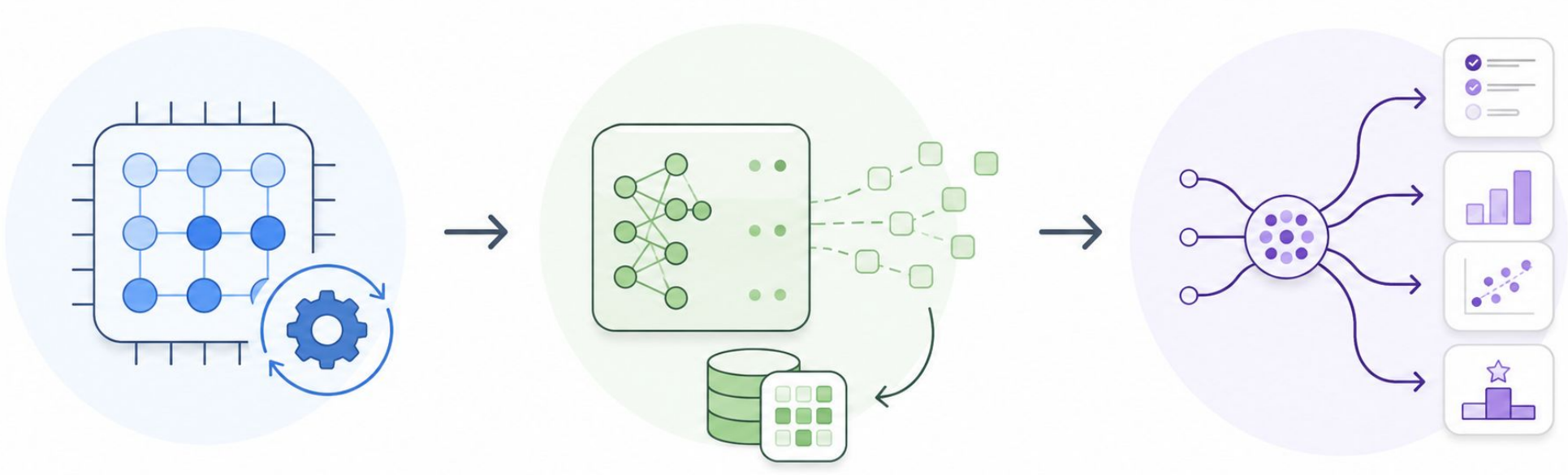
Inference

How can we *serve* foundation models faster under memory constraints?

Evaluation

How can we *capture* LLM capabilities more efficiently and accurately?

Efficiency Across the Foundation-Model Lifecycle



Training

How can we *train* foundation models in fewer optimization steps?

Our solution: *DynMuon* reduces training steps by up to **22% while consistently lowering validation loss over Muon**.

Inference

How can we *serve* foundation models faster under memory constraints?

Our solution: *Randomized KV eviction and learning-based routing* **reduce serving latency by up to 11.96× over SGLang.**

Evaluation

How can we *capture* LLM capabilities more efficiently and accurately?

Our solution: *ECC* calibrates semantic clustering with limited pairwise comparisons, **enabling accurate query-specific LLM capability inference.**

Muon Is a Special Case of Spectral Shaping

- **Usual matrix update**: $M = U\Sigma V^\top$
  - M is typically the momentum-averaged gradient matrix.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Muon Is a Special Case of Spectral Shaping

- **Usual matrix update**: $M = U\Sigma V^\top$ (by SVD)
  - *M* is typically the momentum-averaged gradient matrix.

- **Muon update**: $M = U\Sigma^0 V^\top = UV^\top$
  - flattens singular values; preserves singular directions.
  - achieves strong empirical performance.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Muon Is a Special Case of Spectral Shaping

- **Usual matrix update**: $M = U\Sigma V^\top$ (by SVD)
  - *M* is typically the momentum-averaged gradient matrix.

- **Muon update**: $M = U\Sigma^0 V^\top = UV^\top$
  - flattens singular values; preserves singular directions.
  - achieves strong empirical performance.

- **General view of "spectral shaping"**:
  - adjust how much update strength is assigned to each singular direction.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

From Muon to Power-Law Spectral Shaping

- A natural question:

Should this spectral shaping stay fixed throughout training?

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

From Muon to Power-Law Spectral Shaping

- A natural question:

Should this spectral shaping stay fixed throughout training?

- We study a simple power-law family of spectral shaping:

$$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p),$$

  - $p = 1$: SGD-style update
  - $p = 0$: Muon update
  - $p = -1$: Newton-like inverse-spectrum reweighting

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Training Stages Prefer Different Spectral Shaping

- **Key observation:**
  - The **residual training signal (i.e., remaining distance left to optimum)** is not uniformly distributed across spectral directions.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Training Stages Prefer Different Spectral Shaping

- **Key observation:**
  - The residual training signal (remaining distance left to optimum) is not uniformly distributed across spectral directions.
- **Early training:**
  - Residual signal is concentrated in strong spectral directions (i.e., directions with large singular values).

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Training Stages Prefer Different Spectral Shaping

- **Key observation:**
  - The residual training signal (remaining distance left to optimum) is not uniformly distributed across spectral directions.
- **Early training:**
  - Residual signal is concentrated in strong spectral directions (i.e., directions with large singular values).
- **Late training:**
  - The distribution of residual signal shifts toward weaker directions.
- **Implication:**
  - A fixed spectral weighting is unlikely to be optimal throughout training.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Late Training Benefits from Mildly Negative p

$$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p)$$

- The spectral exponent p controls the relative weighting of spectral directions.
  - **Positive p** emphasizes **strong spectral directions**.
  - **Mildly negative p** emphasizes **weaker spectral directions**.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Late Training Benefits from Mildly Negative p

$$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p)$$

- The spectral exponent p controls the relative weighting of spectral directions.
  - **Positive p** emphasizes **strong spectral directions**.
  - **Mildly negative p** emphasizes **weaker spectral directions**.
  - **Overly negative p** may **amplify *noise*** along weaker directions.

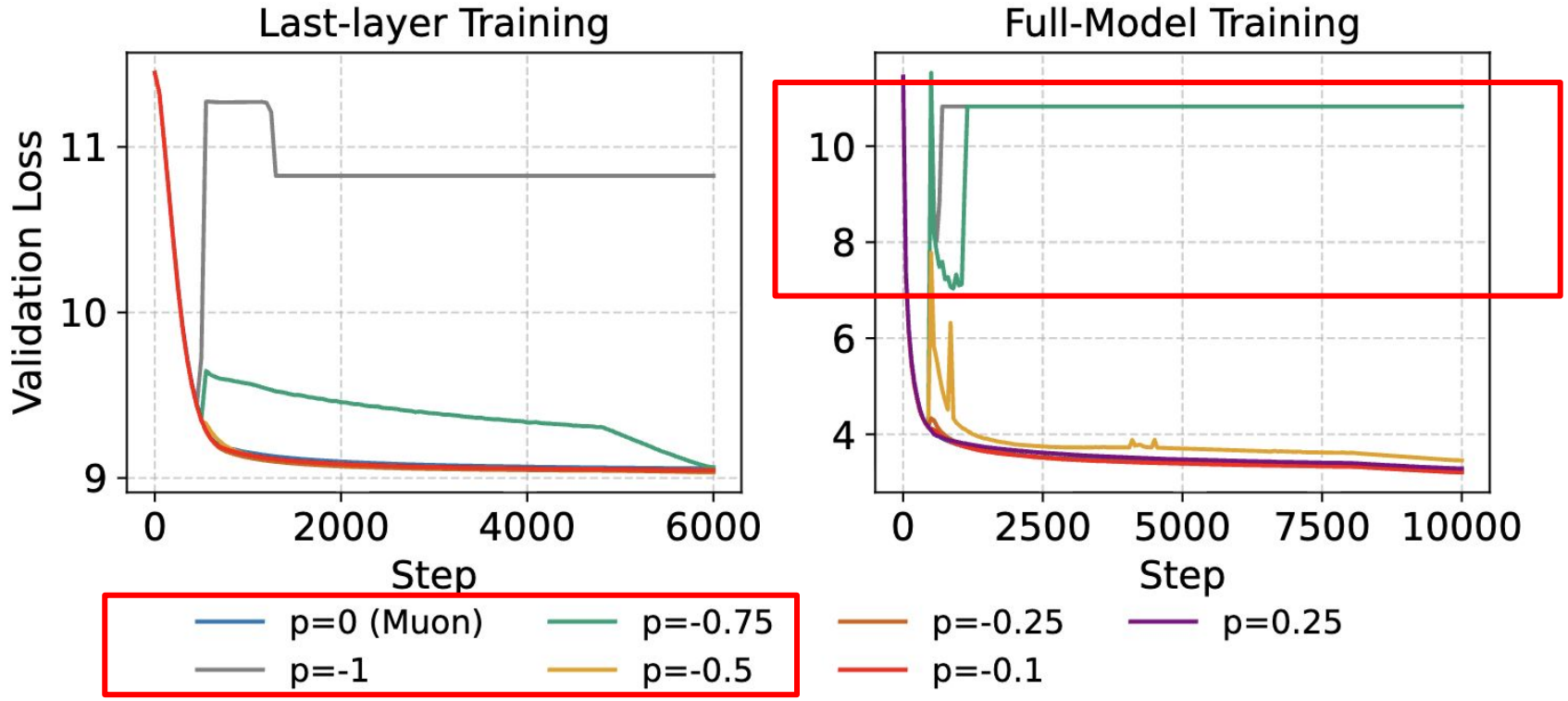
Residual signal shifts → In the late-stage training, p should be mildly negative!

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Late Training Benefits from Mildly Negative p

$$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p)$$

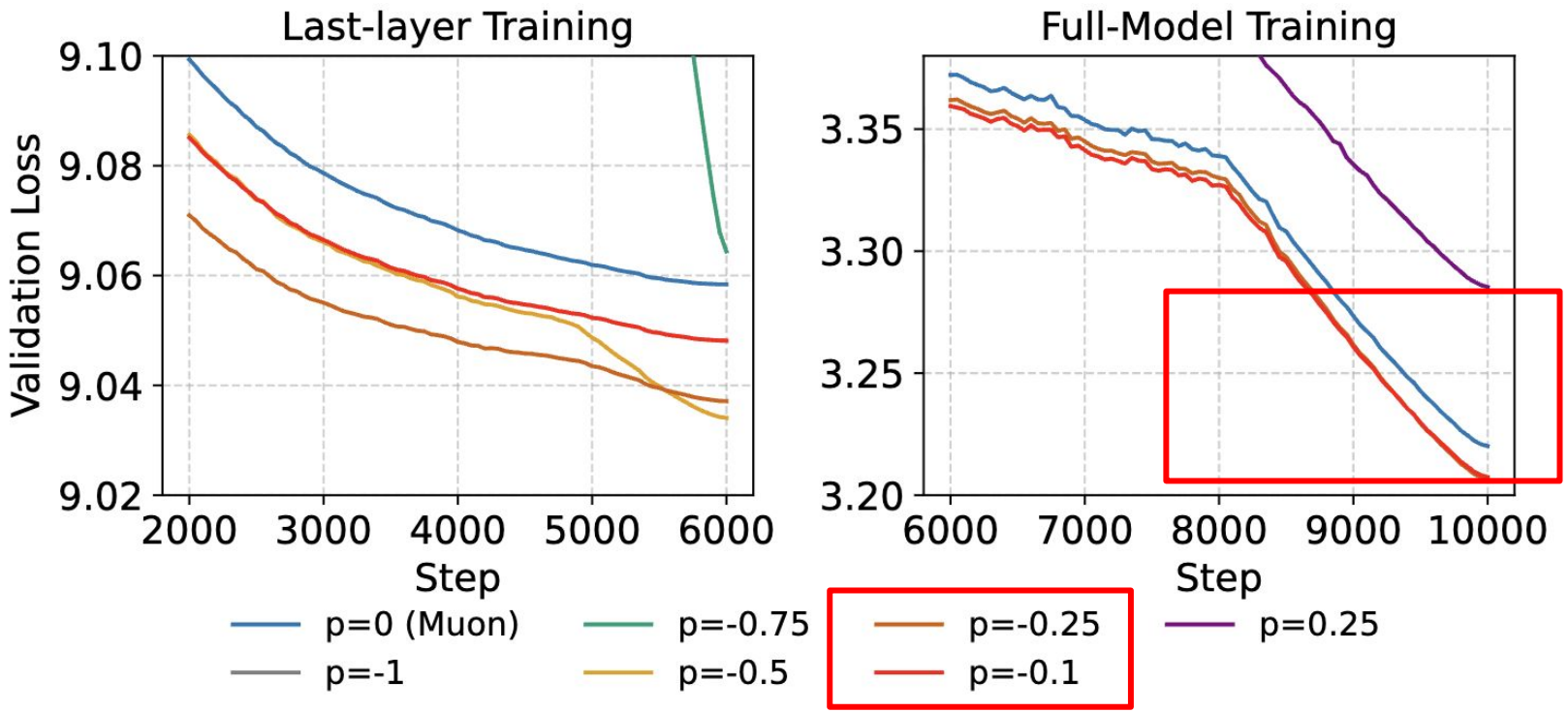
- In late-stage training, p should become **mildly negative!**
- We empirically validate this prediction by changing p to different values at step 500.



Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Late Training Benefits from Mildly Negative p

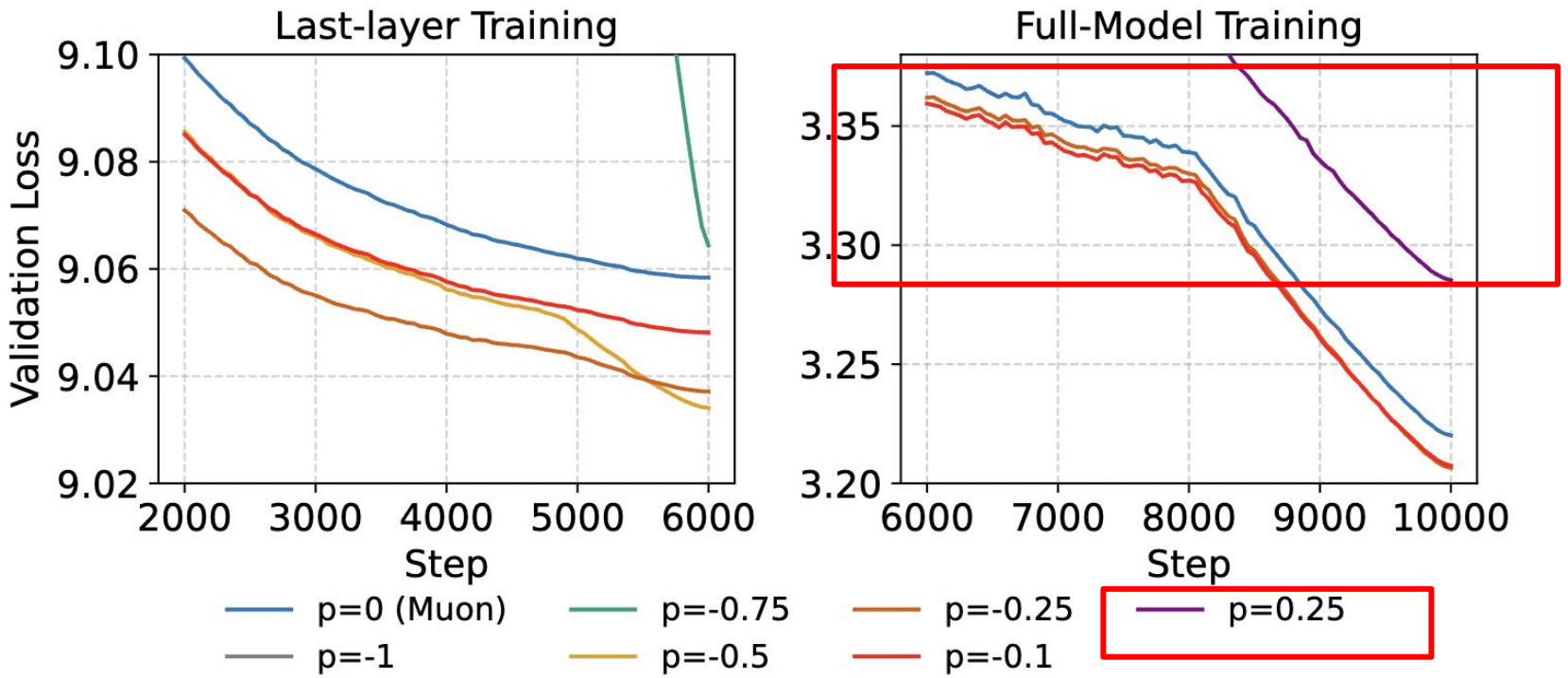
- $$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p)$$
- In late-stage training, p should become **mildly negative!**
- We empirically validate this prediction by changing p to different values at step 500.



Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Late Training Benefits from Mildly Negative p

- $$D^{(p)} := U\Sigma^p V^\top, \qquad \Sigma^p = \mathrm{diag}(\sigma_1^p, \ldots, \sigma_r^p)$$
- In late-stage training, p should become **mildly negative!**
- We empirically validate this prediction by changing p to different values at step 500.



Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Early Training Benefits from Positive p

- **Early-stage intuition**:
  Before residual signal shifts toward weaker directions, dominant directions still carry useful signal.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

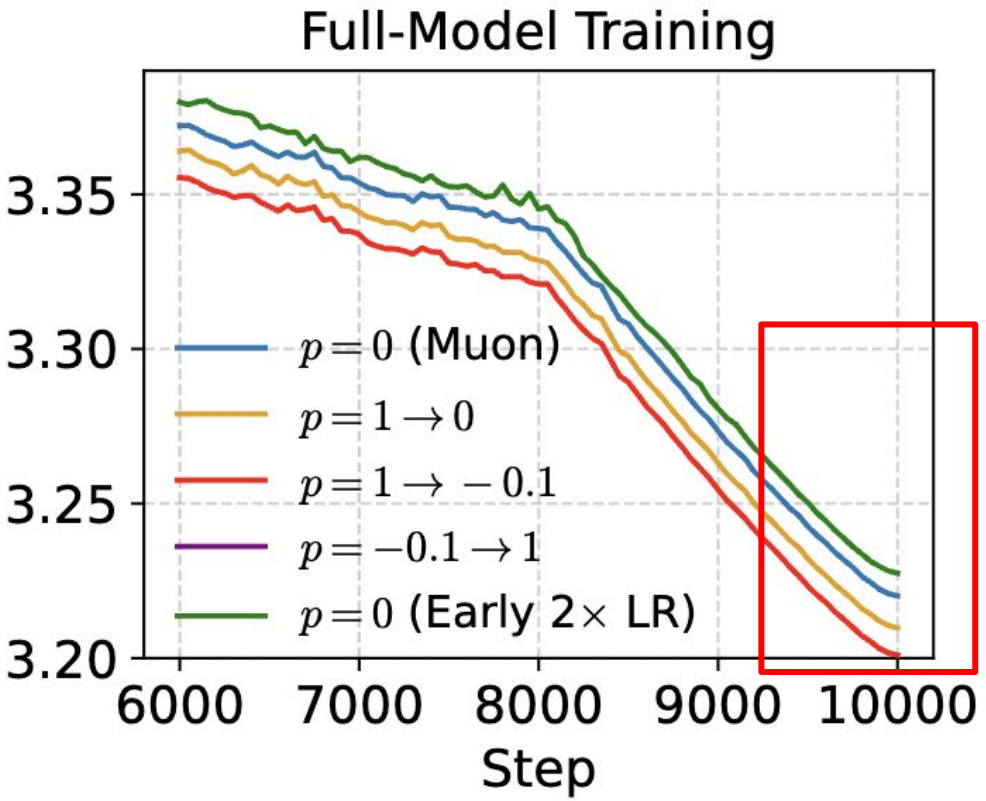
Early Training Benefits from Positive p

- **Early-stage intuition**:
  Before residual signal shifts toward weaker directions, dominant directions still carry useful signal.
- A positive p can:
  - prioritizes dominant spectral directions
  - accelerates early-stage signal reduction

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

Early Training Benefits from Positive p

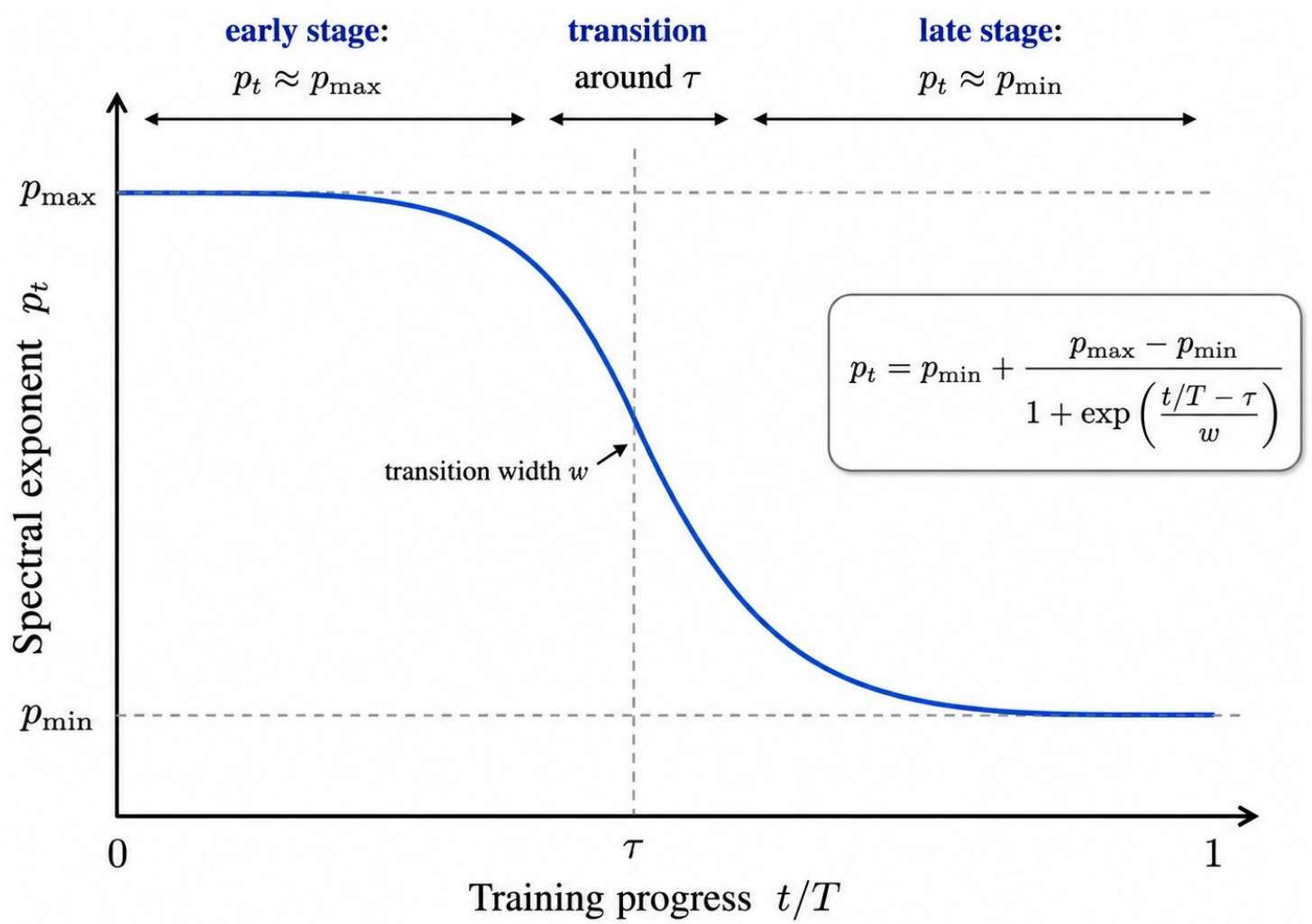
- **Early-stage intuition**:
  Before residual signal shifts toward weaker directions, dominant directions still carry useful signal.
- A positive p can:
  - prioritizes dominant spectral directions
  - accelerates early-stage signal reduction
- Empirical validation:
  Set $p = 1$ in the first 500 steps, then switch to $p = 0$ or mildly negative $p = -0.1$.



Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

DynMuon: Dynamic Spectral Shaping

- DynMuon uses a simple logistic schedule to decrease p over training via a simple logistic schedule.



DynMuon: Dynamic Spectral Shaping

- Exact SVD can implement $U\Sigma^p V^\top$ for any fractional p, but it is too expensive.

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

DynMuon: Dynamic Spectral Shaping

- Exact SVD can implement $U\Sigma^p V^\top$ for any fractional p, but it is too expensive.
- **Key idea**: reuse the fast Muon update and add a lightweight spectral correction.
- This keeps DynMuon close to Muon's efficiency while allowing p to change across training.

$$\text{DynMuon update} \approx \underbrace{\text{Muon update}}_{\text{fast}} + \underbrace{\text{spectral correction}}_{\text{lightweight}}$$

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

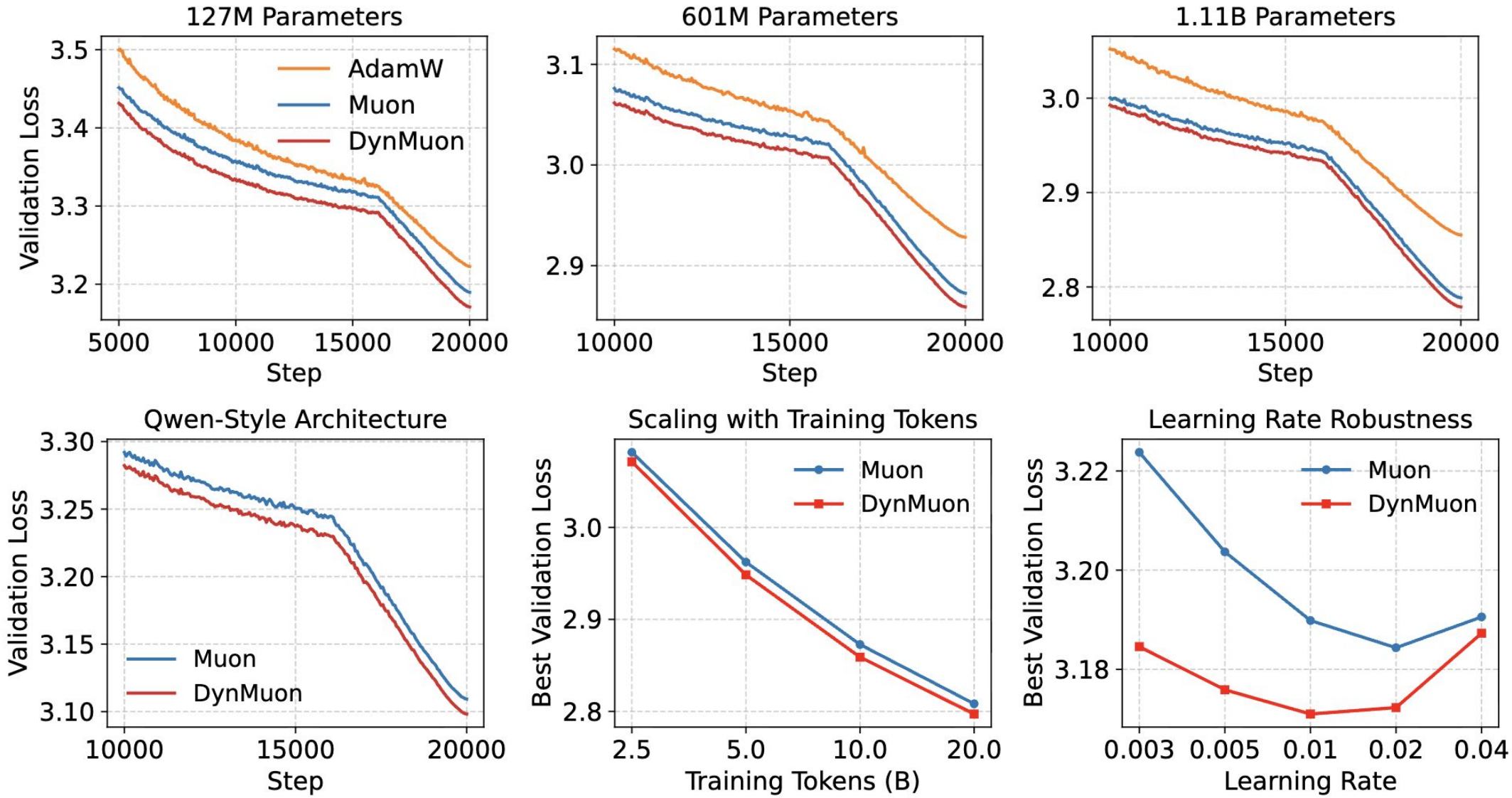
Results

Table 1: Performance and efficiency of DynMuon relative to Muon across GPT-style model scales. Steps to Target uses the validation loss reached by Muon at 80% of training as the target. Step Saving reports the relative step reduction, and Per-Step Time is the average ms/step.

Tokens	Method (Size)	Best Val. Loss (↓)	Steps to Target (↓)	Step Saving (↑)	Per-Step Time (ms)
10B	Muon (127M)	3.190	16000	0.0%	1142.4
	DynMuon (127M)	**3.171**	**12500**	**21.9%**	1150.3
	Muon (601M)	2.872	16000	0.0%	4121.7
	DynMuon (601M)	**2.858**	**13950**	**12.8%**	4200.1
	Muon (1.1B)	2.788	16000	0.0%	6883.3
	DynMuon (1.1B)	**2.776**	**14300**	**10.6%**	7055.8
20B	Muon (127M)	3.139	30400	0.0%	1137.3
	DynMuon (127M)	**3.124**	**22350**	**26.5%**	1151.8
	Muon (601M)	2.808	30400	0.0%	4126.2
	DynMuon (601M)	**2.797**	**25000**	**17.8%**	4184.8
	Muon (1.1B)	2.722	30400	0.0%	6889.77
	DynMuon (1.1B)	**2.713**	**26450**	**13.0%**	6910.1

Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

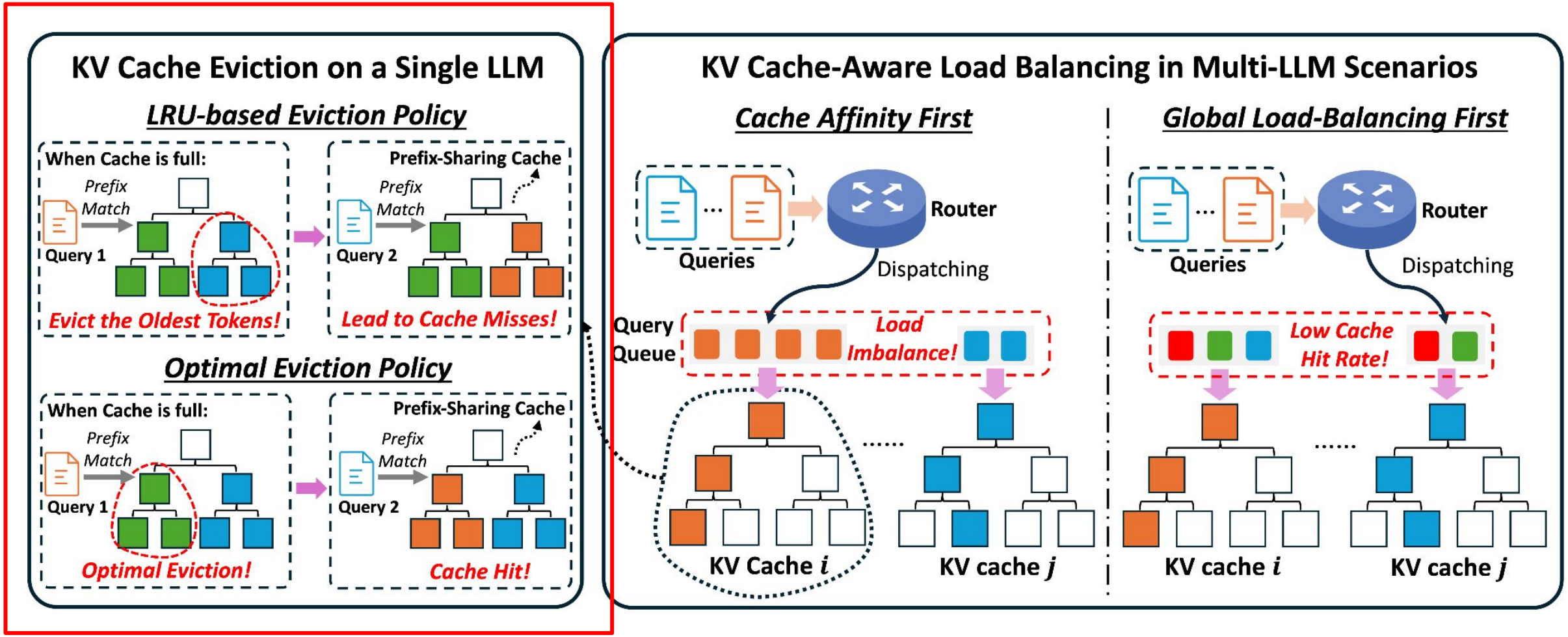
Results



Wu, F., Shah, R., Silwal, S. and Zhang, Q., DynMuon: A Dynamic Spectral Shaping View of Muon, arXiv:2605.17109, 2026.

KV Cache-Aware Load Balancing

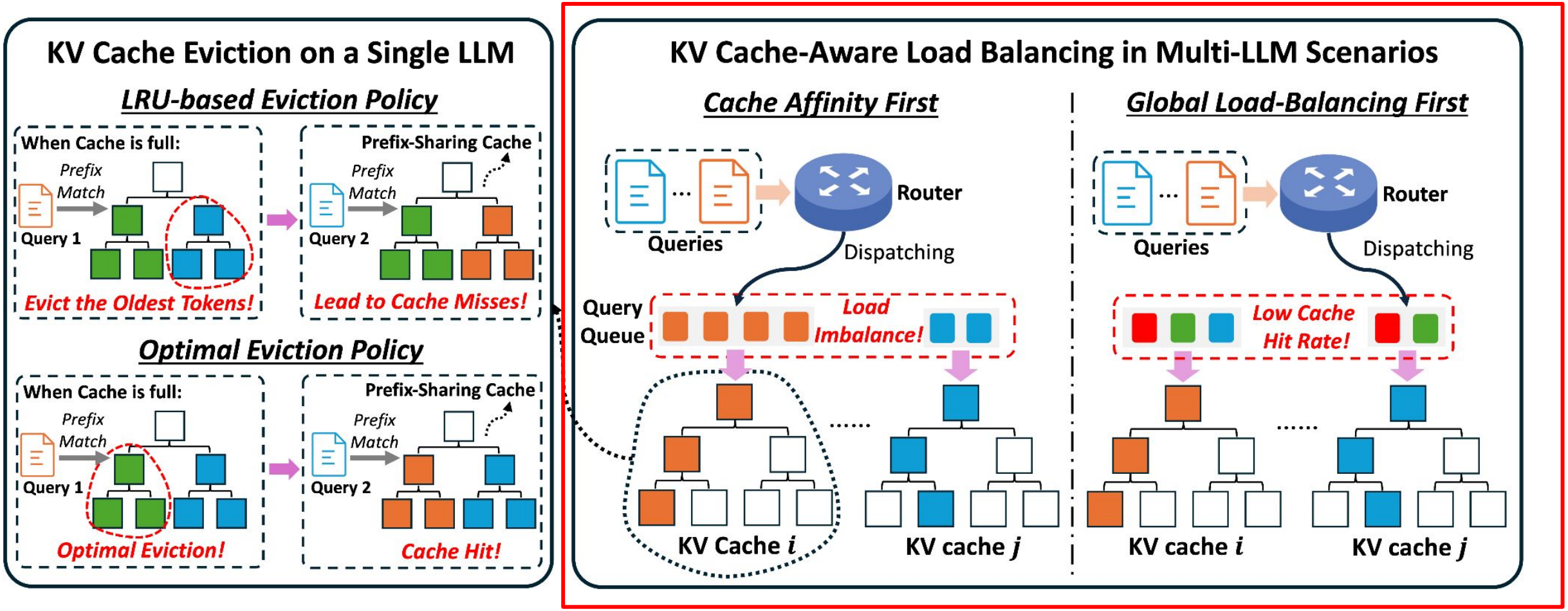
- Core challenge in multi-LLM serving: cache reuse and global load balancing can conflict.



Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

KV Cache-Aware Load Balancing

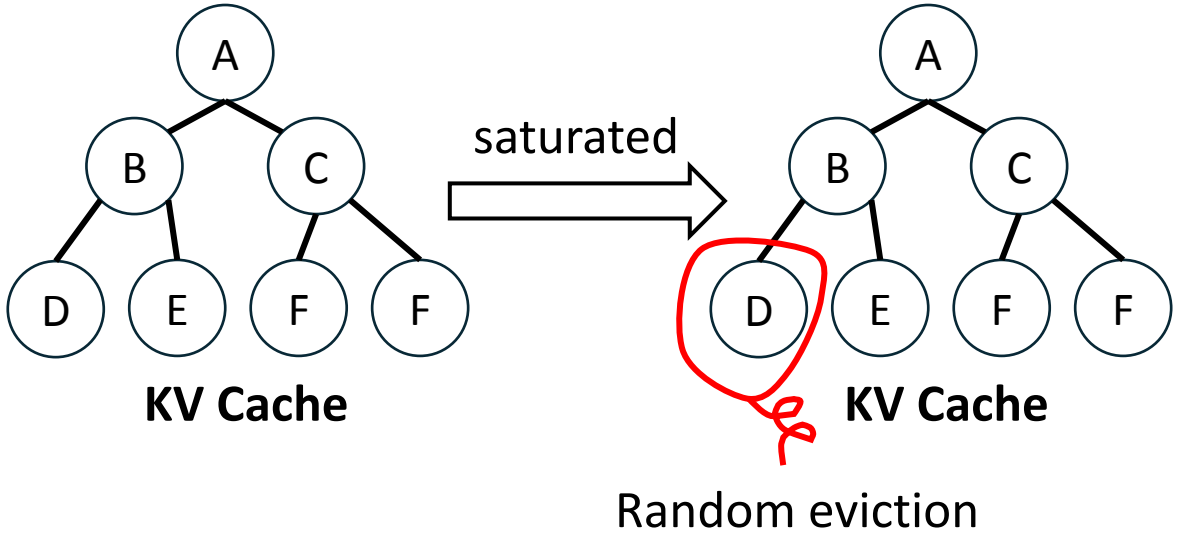
- Core challenge in multi-LLM serving: cache reuse and global load balancing can conflict.



Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Randomized **L**eaf **T**oken eviction (RLT)

- **Key idea**: Replace **deterministic** LRU eviction with **randomized** leaf-token eviction.
- **Initution**: Randomization provides robustness against complex and dynamic query workloads.
- When saturated, RLT **randomly evicts an available leaf token**.



Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Learning-Based Routing for Multi-LLM Serving

- Existing systems such as **NVIDIA Dynamo** and **SGLang** rely on fixed routing heuristics based on cache hit rate and queue length.

- Limitation:
  - Inherently **static** and unable to adapt to dynamic query arrival patterns.

Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Learning-Based Routing for Multi-LLM Serving

- Existing systems such as **NVIDIA Dynamo** and **SGLang** rely on fixed routing heuristics based on cache hit rate and queue length.

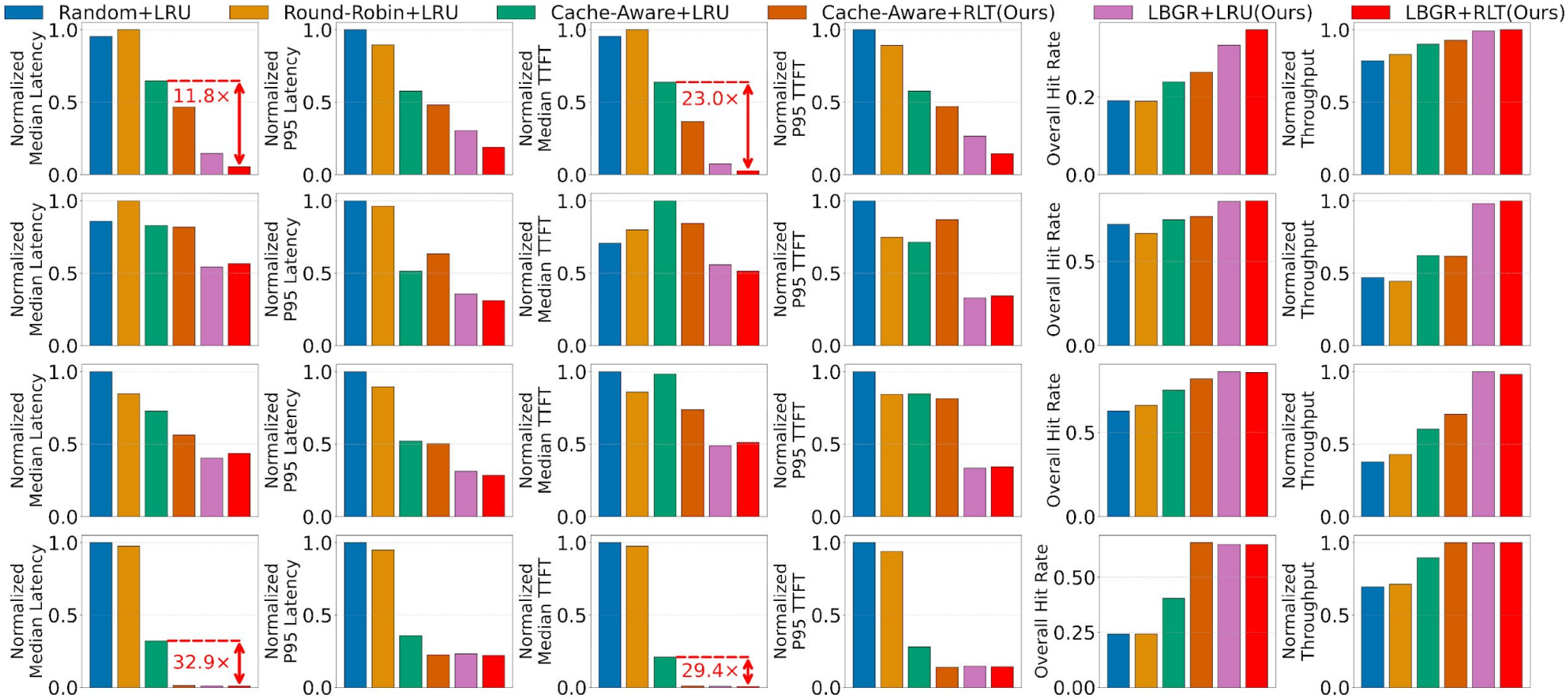
- Limitation:
  - Inherently **static** and unable to adapt to dynamic query arrival patterns.

- **Our approach:** we propose **LBGR**, a lightweight **learning-based routing** algorithm
  - learns an online latency predictor and routes each query to the model with the lowest estimated latency.
    - Estimates KV cache hit rate for each model
    - Accounts for the queue load for each model

Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Main Results

- Extensive experiments demonstrate improvements of up to **6.92× in cache hit rate,11.96× reduction in latency, 14.06× reduction in time-to-first-token (TTFT), and 77.4% increase in throughput** compared with the **state-of-the-art baseline, SGLang.**



Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Main Results

- Introduce negligible overhead compared with baselines.

Table 2: Ablation comparison of performance and runtime overhead between Cache-Aware+LRU and our methods on the GSP benchmark. Time-based metrics (↓) are reported in milliseconds (ms), while hit rate (↑) and throughput (↑) are measured in percentage and requests per second, respectively.

Method	P50 Latency	P95 Latency	P50 TTFT	P95 TTFT	Hit Rate	Throughput	Average Eviction Time	Average Routing Time
Cache-Aware+LRU	26680.55	46766.77	25022.76	46139.36	23.89%	10.73	0.13	**0.47**
Cache-Aware+RLT (Ours)	19191.25	38917.27	14332.81	37504.69	26.36%	11.05	0.71	0.51
LBGR+LRU (Ours)	6025.11	24561.47	2958.01	21073.78	33.33%	11.80	**0.09**	1.03
LBGR+RLT (Ours)	**2263.61**	**15334.89**	**1088.57**	**11495.05**	**37.31%**	**11.92**	1.05	1.45

Wu, F., Silwal, S. and Zhang, Q., Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective, in ICLR 2026

Large Language Model with Diverse Capabilities

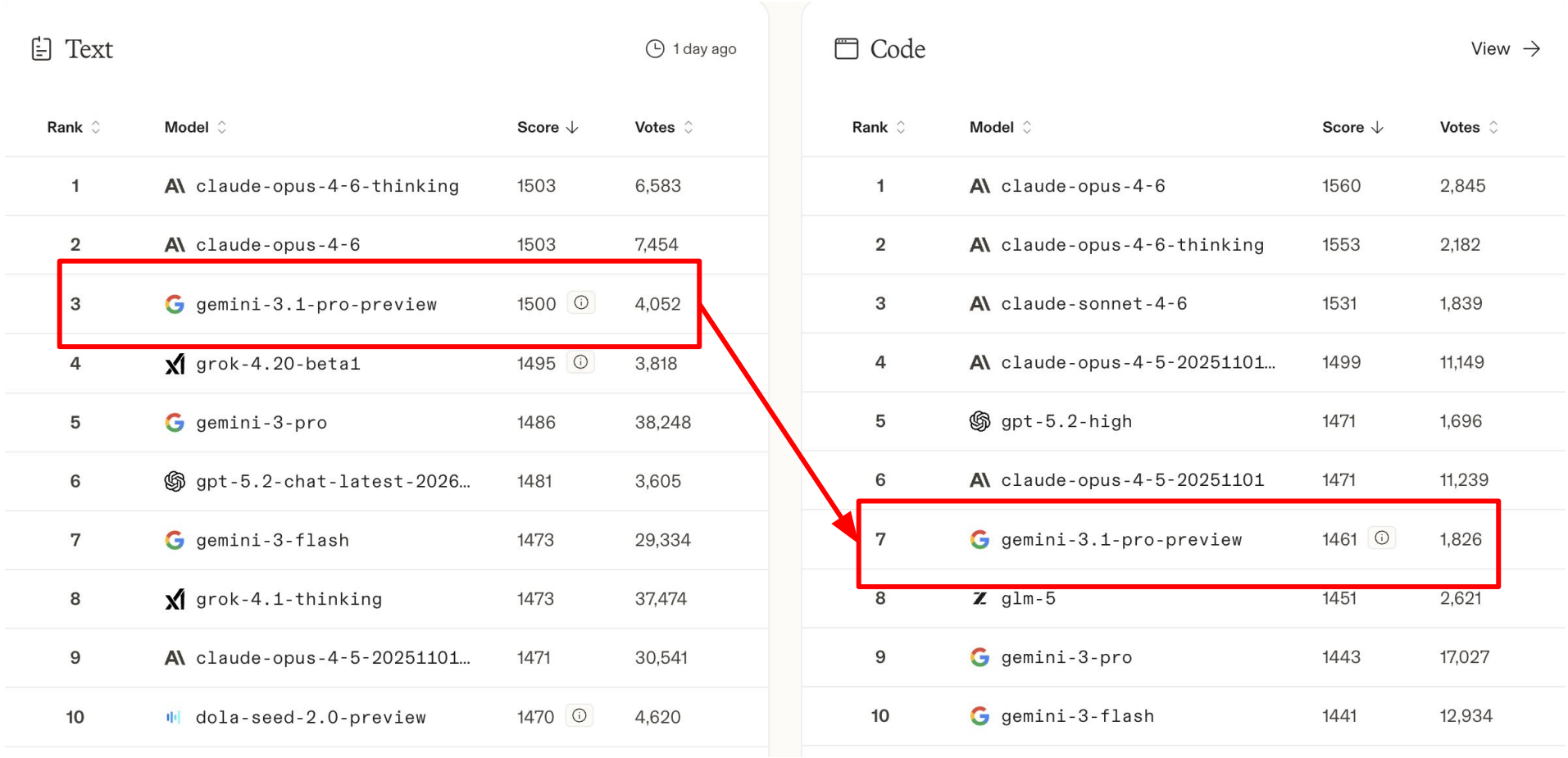
- LLMs trained on extensive corpora exhibit diverse capabilities (e.g., math, coding).

Benchmark		Gemini 3.1 Pro Thinking (High)	Gemini 3 Pro Thinking (High)	Sonnet 4.6 Thinking (Max)	Opus 4.6 Thinking (Max)	GPT-5.2 Thinking (xhigh)	GPT-5.3-Codex Thinking (xhigh)
Humanity's Last Exam Academic reasoning (full set, text + MM)	No tools	**44.4%**	37.5%	33.2%	40.0%	34.5%	—
	Search (blocklist) + Code	51.4%	45.8%	49.0%	**53.1%**	45.5%	—
ARC-AGI-2 Abstract reasoning puzzles	ARC Prize Verified	**77.1%**	31.1%	58.3%	68.8%	52.9%	—
GPQA Diamond Scientific knowledge	No tools	**94.3%**	91.9%	89.9%	91.3%	92.4%	—
Terminal-Bench 2.0 Agentic terminal coding	Terminus-2 harness	**68.5%**	56.9%	59.1%	65.4%	54.0%	64.7%
	Other best self-reported harness	—	—	—	—	62.2% (Codex)	**77.3%** (Codex)
SWE-Bench Verified Agentic coding	Single attempt	80.6%	76.2%	79.6%	**80.8%**	80.0%	—
SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	54.2%	43.3%	—	—	55.6%	**56.8%**

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Large Language Model with Diverse Capabilities

- LLMs trained on extensive corpora exhibit diverse capabilities (e.g., math, coding).
- **Different LLMs excel in different tasks.**



Text — 1 day ago

Rank	Model	Score	Votes
1	claude-opus-4-6-thinking	1503	6,583
2	claude-opus-4-6	1503	7,454
3	gemini-3.1-pro-preview	1500	4,052
4	grok-4.20-beta1	1495	3,818
5	gemini-3-pro	1486	38,248
6	gpt-5.2-chat-latest-2026...	1481	3,605
7	gemini-3-flash	1473	29,334
8	grok-4.1-thinking	1473	37,474
9	claude-opus-4-5-20251101...	1471	30,541
10	dola-seed-2.0-preview	1470	4,620

Code — View →

Rank	Model	Score	Votes
1	claude-opus-4-6	1560	2,845
2	claude-opus-4-6-thinking	1553	2,182
3	claude-sonnet-4-6	1531	1,839
4	claude-opus-4-5-20251101...	1499	11,149
5	gpt-5.2-high	1471	1,696
6	claude-opus-4-5-20251101	1471	11,239
7	gemini-3.1-pro-preview	1461	1,826
8	glm-5	1451	2,621
9	gemini-3-pro	1443	17,027
10	gemini-3-flash	1441	12,934

Large Language Model with Diverse Capabilities

- One fundamental question for LLM-based applications:

How can we accurately capture and assess their strengths?

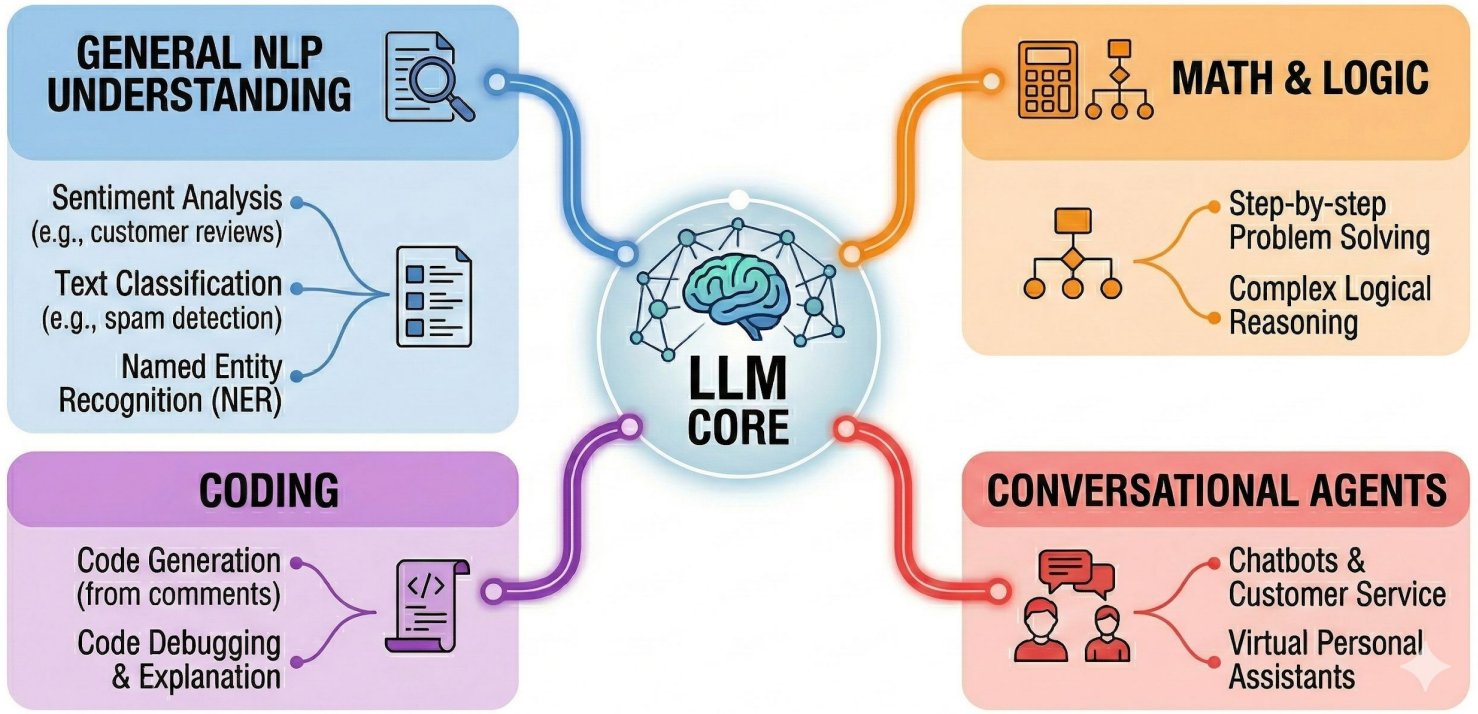
Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Large Language Model with Diverse Capabilities

- One fundamental question for LLM-based applications:

  How can we accurately capture and assess their strengths?

- One widely adopted method: Evaluating them on query benchmarks!



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Evaluating LLM Capabilities via Queries

- Queries implicitly specify diverse capability requirements.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

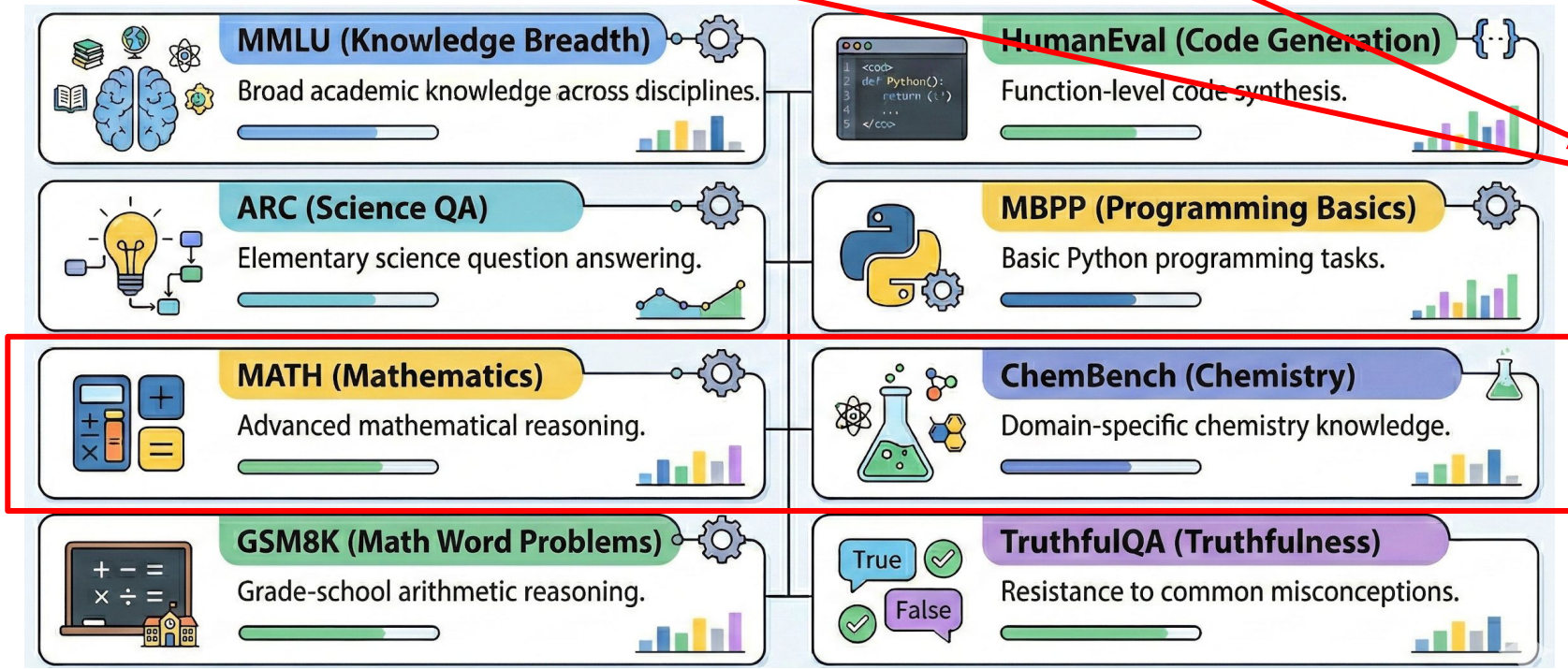
Evaluating LLM Capabilities via Queries

- Queries implicitly specify diverse capability requirements.
- LLM performance (capability) is inherently query-dependent.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Evaluating LLM Capabilities via Queries

- Queries implicitly specify diverse capability requirements.
- LLM performance (capability) is inherently query-dependent.
- **Common practice: human semantic labels → capability proxies → subset evaluation**



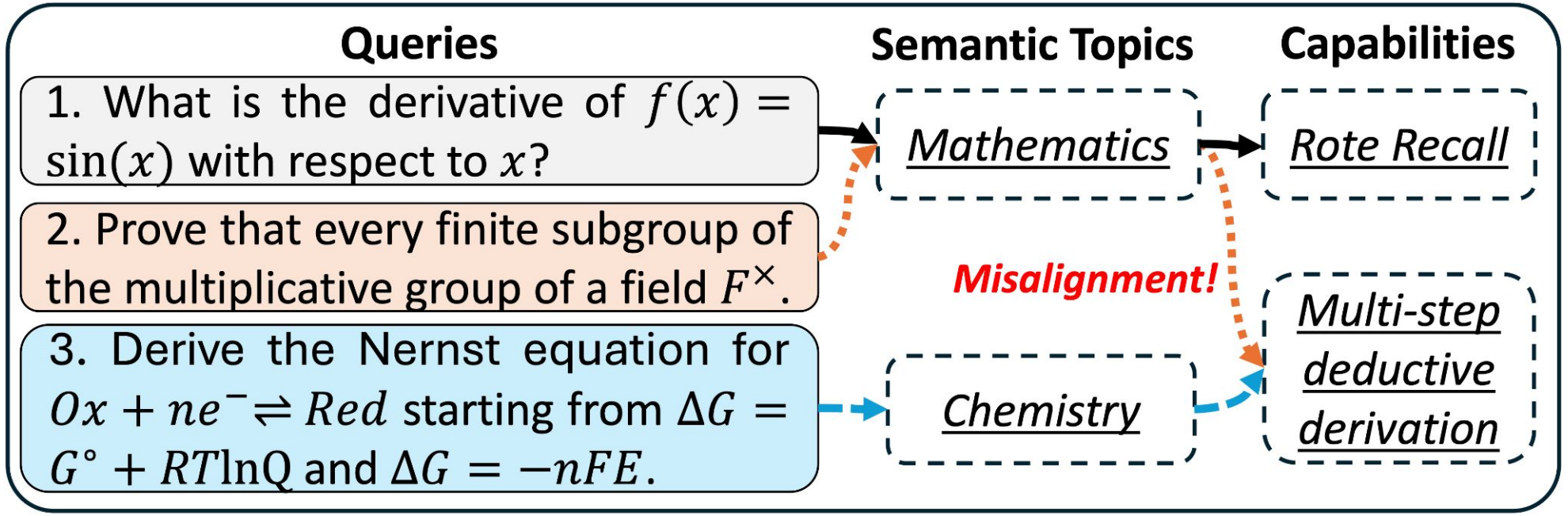
Do they align with each other?

Subsets based on different semantics

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Semantic Labels Do Not Always Reflect Underlying Capabilities!

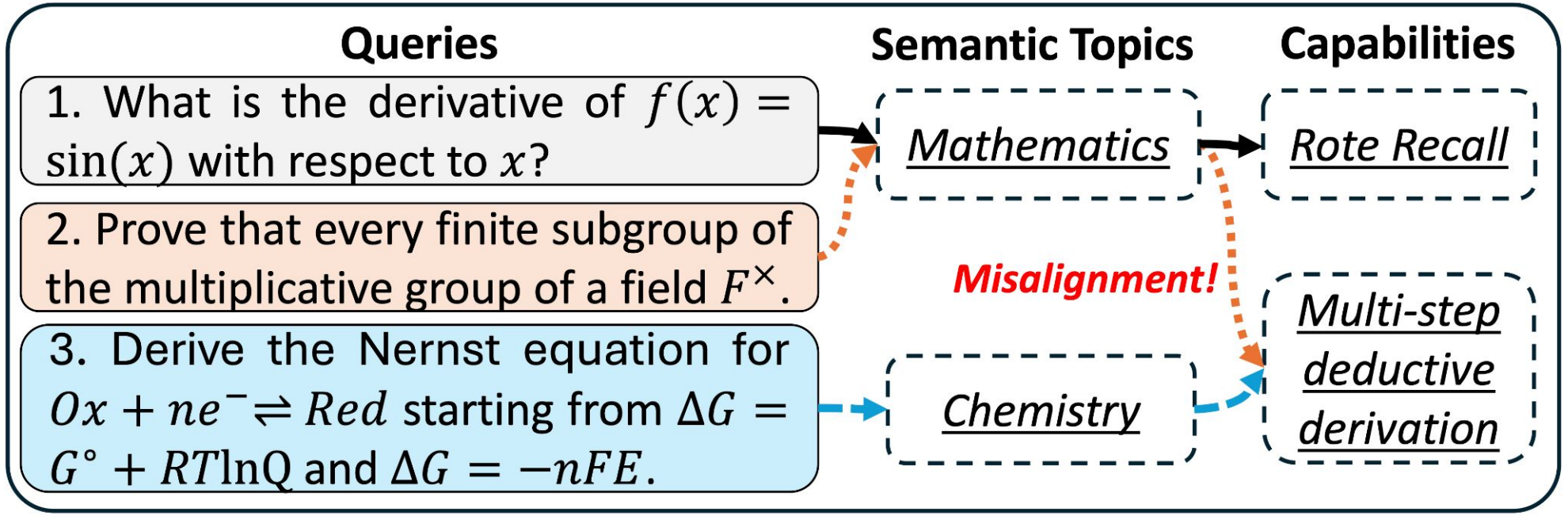
- A single “Mathematics” label can include queries requiring vastly different levels of capabilities, or even a combination of distinct capability requirements.



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Semantic Labels Do Not Always Reflect Underlying Capabilities!

- A single “Mathematics” label can include queries requiring vastly different levels of capabilities, or even a combination of distinct capability requirements.
- Conversely, queries that require the similar underlying capabilities may be scattered across different semantic subsets due to superficial topical differences.



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Semantic Labels Do Not Always Reflect Underlying Capabilities!

- This (human-labeled) misalignment with the true capability distributions can limit, or even degrade, the generalizability of model capability estimates to unseen queries.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Semantic Labels Do Not Always Reflect Underlying Capabilities!

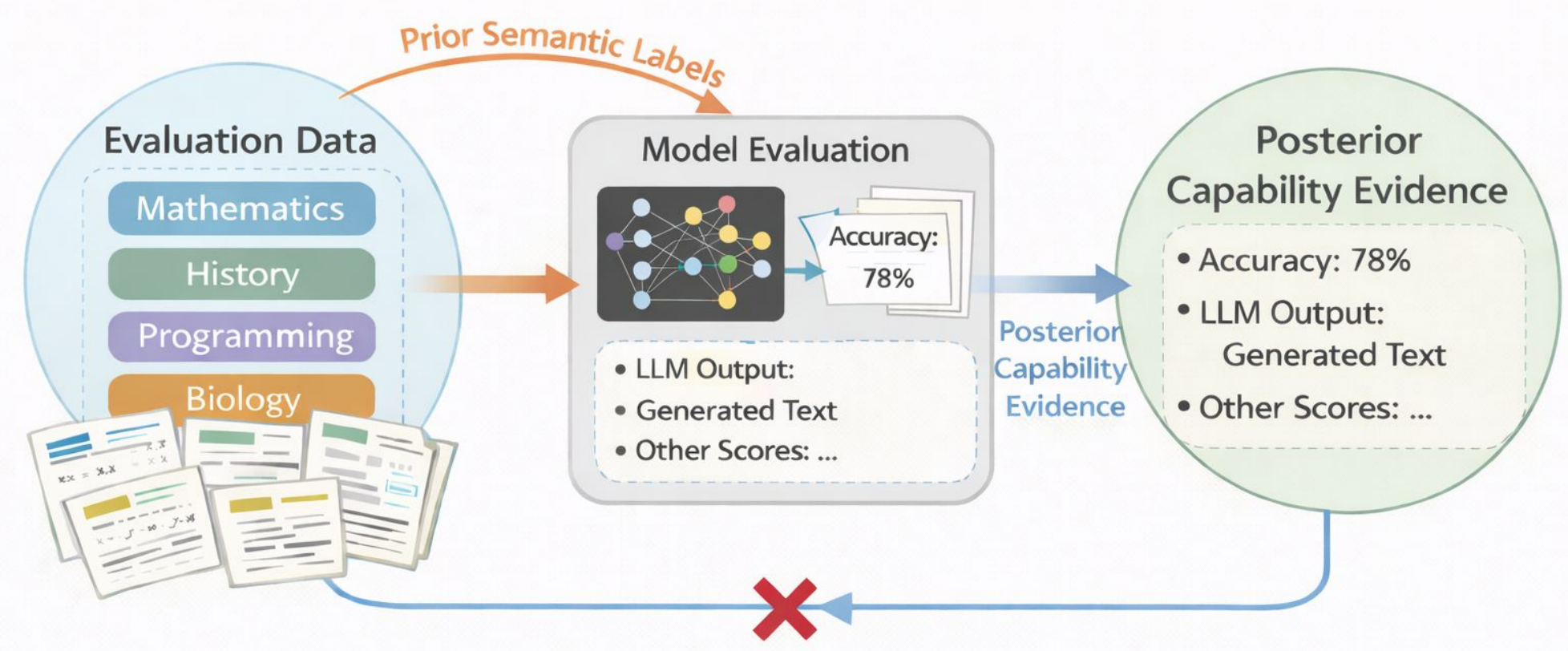
- This (human-labeled) misalignment with the true capability distributions can limit, or even degrade, the generalizability of model capability estimates to unseen queries.

- **Goal**: learn capability-aligned query clusters that better capture underlying capability structure.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

From Semantics to Capability Evidence

- Key observation: Evaluation data is split and organized only by ***prior*** semantics, while ***posterior*** evidence of model capability is only used ***after model evaluation***.



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Bradley-Terry Model — A Posterior Signal of Model Capability

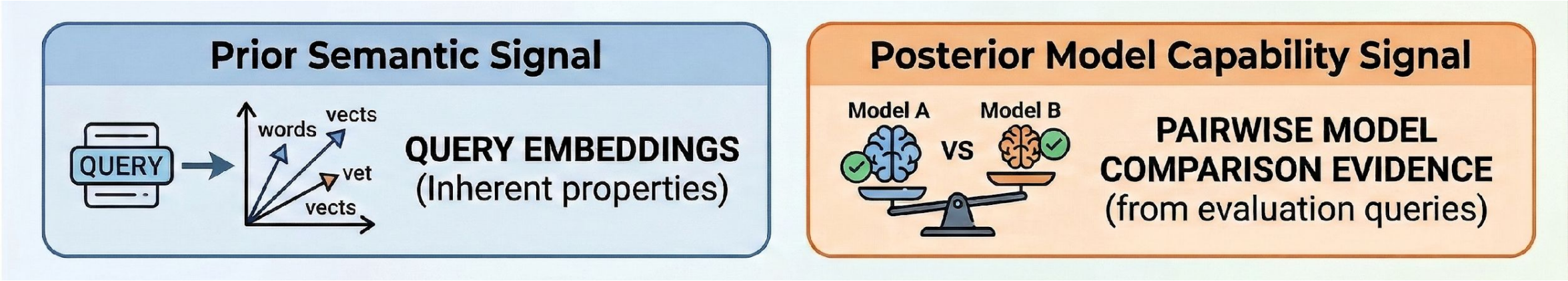
- A Bradley-Terry (BT) model assigns a latent strength score to each item by fitting pairwise comparison outcomes among the items.
  - In our setting, the items are LLMs, and comparisons are between LLM responses to the same query.

- Given limited pairwise comparisons among M LLMs, BT model can estimate a capability ranking $\boldsymbol{\theta} \in \mathbb{R}^M$ for these LLMs.

- This offers an efficient way to encode the LLM capabilities.

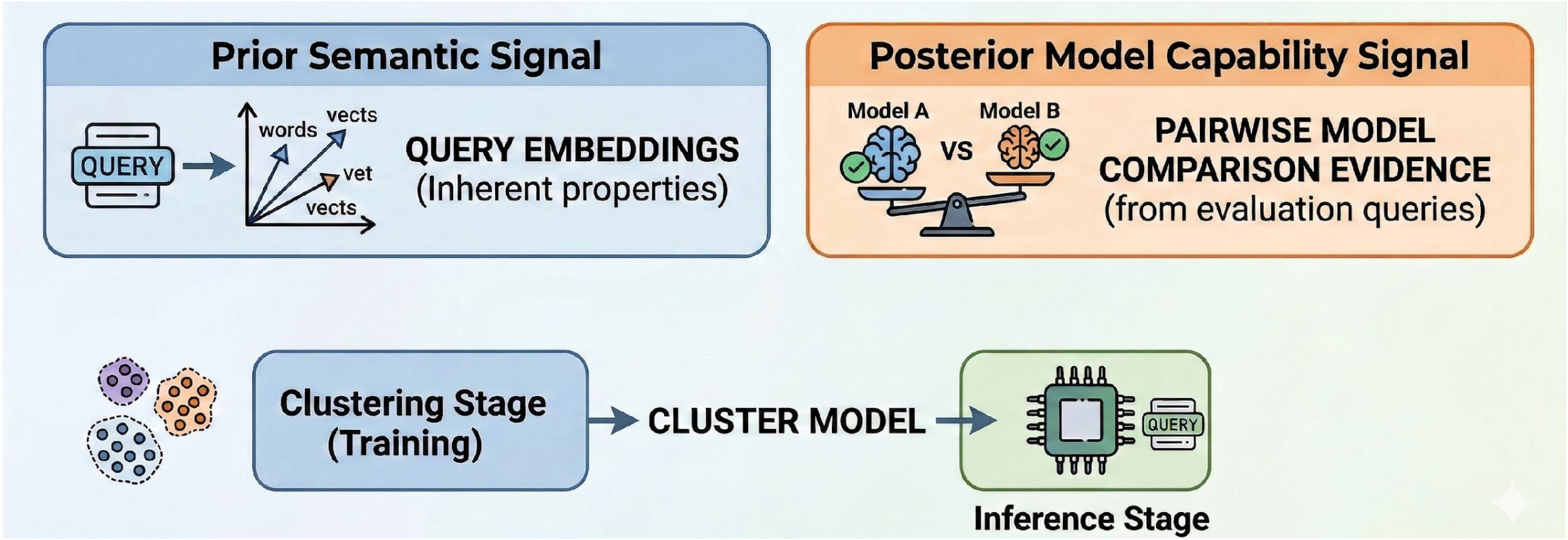
Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Two Types of Signals



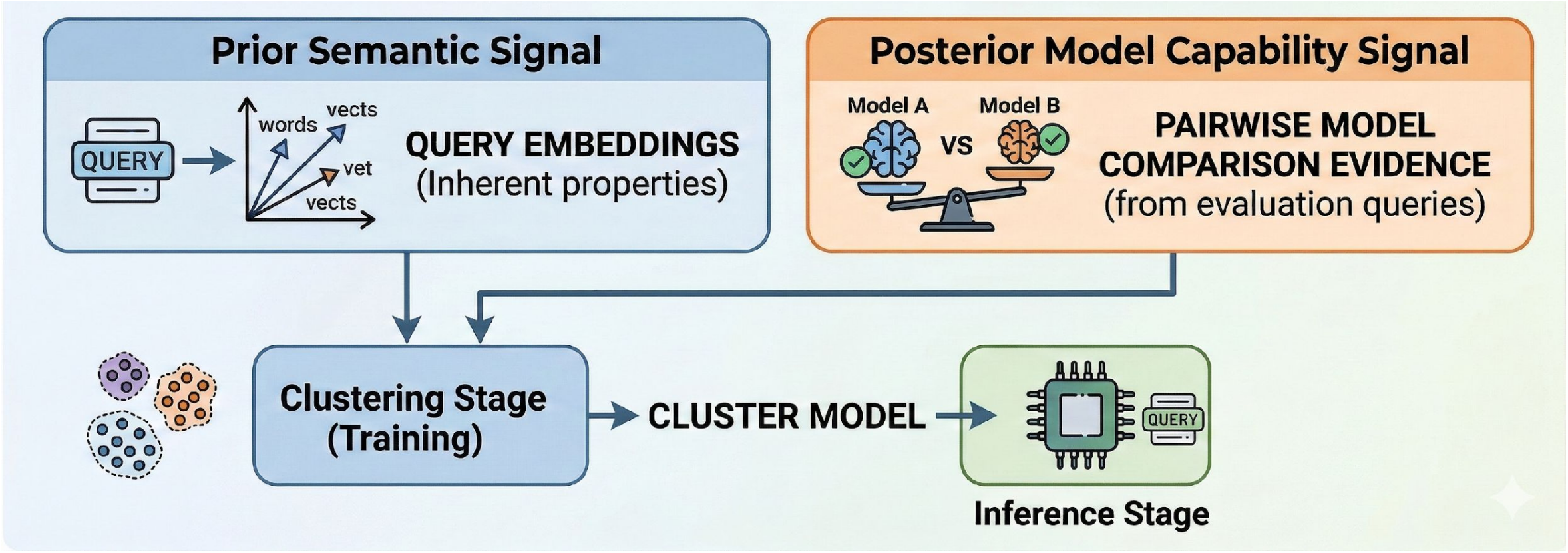
Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Two Stages



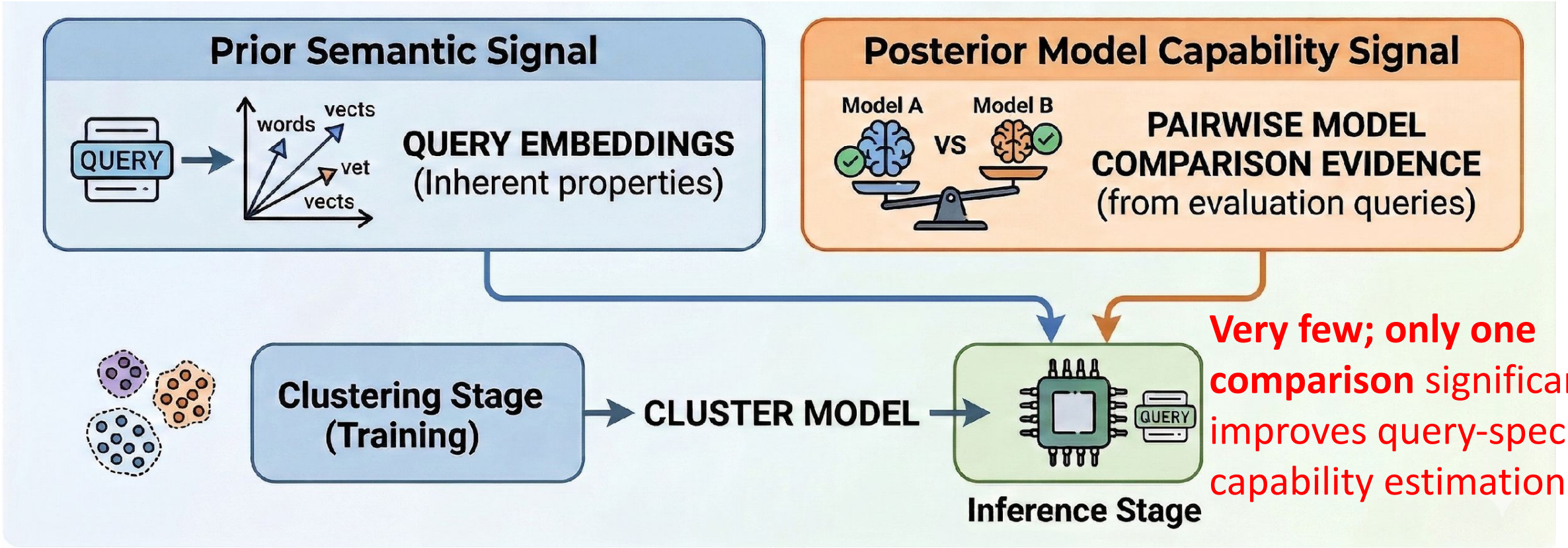
Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Evidence-Calibrated Clustering (ECC)



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Evidence-Calibrated Clustering (ECC)



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Quantitative Results

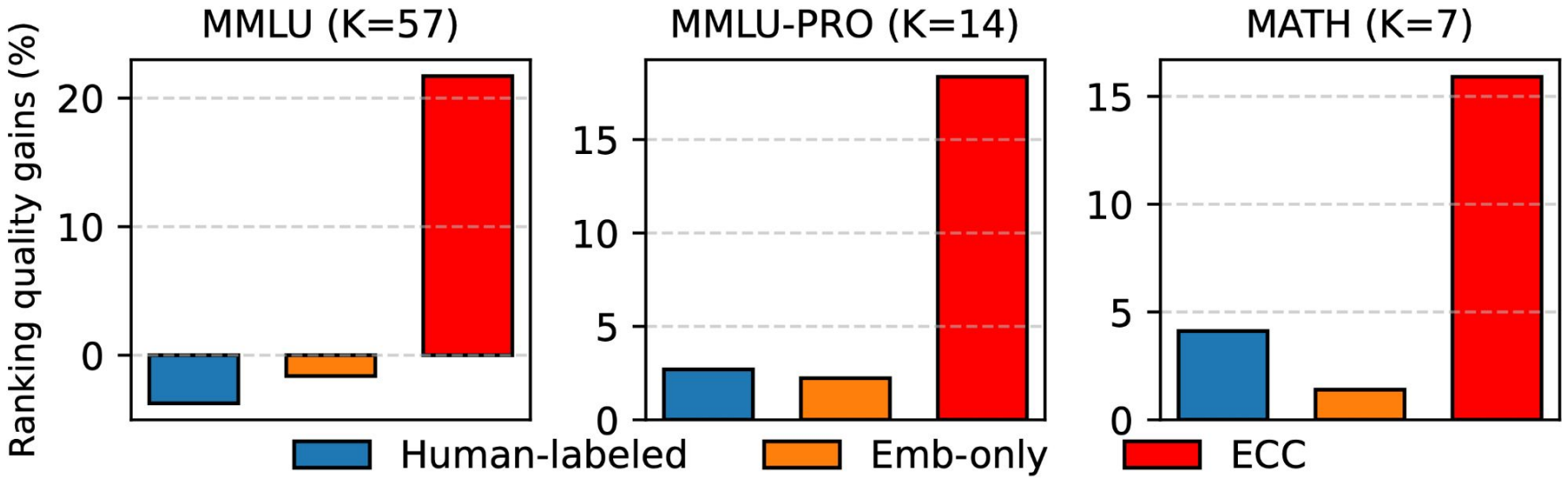
- **Metric**: Ranking quality gains.
  - **average per-query ranking-accuracy gain on unseen queries** relative to a single global BT model without clustering.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Quantitative Results

- **Metric**: Ranking quality gains.
  - **average per-query ranking-accuracy gain on unseen queries** relative to a single global BT model without clustering.

- **RQ**: How does ECC compare to human-labeled/embedding-only data clustering?



Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

ECC Reveals Capability Structure Beyond Embeddings

- We conduct a qualitative case study comparing ECC with embedding-only clustering (semantic-based clustering) and observe two patterns.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

ECC Reveals Capability Structure Beyond Embeddings

- We conduct a qualitative case study comparing ECC with embedding-only clustering (semantic-based clustering) and observe two patterns.

- Within-Embedding Split
  - For queries that are close in embedding space, ECC ***separates*** them by capability needs.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

ECC Reveals Capability Structure Beyond Embeddings

- We conduct a qualitative case study comparing ECC with embedding-only clustering (semantic-based clustering) and observe two patterns.

- Within-Embedding Split
  - For queries that are close in embedding space, ECC ***separates*** them by capability needs.

- Cross-Embedding Merge
  - For queries with different semantics, ECC ***merges*** them together when they share the same capability requirement.

Wu, F., Silwal, S. and Zhang, Q., Capturing LLM Capabilities via Evidence-Calibrated Query Clustering, arXiv:2605.17110, 2026.

Thanks & Questions?

- Related Papers:
  - **Fangzhou Wu**, Rikhav Shah, Sandeep Silwal, and Qiuyi Zhang. ***DynMuon: A Dynamic Spectral Shaping View of Muon***. arXiv preprint arXiv:2605.17109, 2026.
  - **Fangzhou Wu**, Sandeep Silwal, and Qiuyi Zhang. ***Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective***. ICLR 2026.
  - **Fangzhou Wu**, Sandeep Silwal, and Qiuyi Zhang. ***Capturing LLM Capabilities via Evidence-Calibrated Query Clustering***. arXiv preprint arXiv:2605.17110, 2026.