PaSM: Per-Query Trust-Aware Serving for Merged Open-Weight LLMs

Anonymous ACL submission

Abstract

Open-weight Large Language Models are increasingly composed at deployment time via weight-space merging, which combines released checkpoints into a single served model that inherits complementary capabilities without additional training. Existing recipes commit to one fixed merge for all queries. This commitment is inadequate for multi-objective serving: different objectives peak at different points in the merge space, so any single checkpoint very likely sacrifices at least one of them, and the served decision cannot be inspected or attributed after the fact. Different queries, moreover, prefer different regions of the same space, so the optimal merge is a property of the query distribution rather than of any individual query. We propose **PaSM** (**Pa**reto **S**pecialist **M**ixture of routers), a per-query serving framework for merged LLMs that compiles a finite Pareto bank of merged checkpoints, routes each query through a validation-selected specialist only when held-out evidence supports it, and logs every served decision on a per-query Trust Card. The framework applies to any multi-objective serving setting over merged checkpoints where per-query utility is measurable on a held-out fold. We instantiate it on **trust-sensitive serving**, where truthfulness, safety, and calibration must hold jointly. Across 336 audit cells over four released checkpoint pairs, PaSM lifts mean utility by $+6.74$ pp over the strongest feasible baseline and $+2.17$ pp over the best fixed single-λ baseline, at just 1.06 ms median cached latency.

1 Introduction

Open-weight Large Language Models are increasingly deployed via weight-space merging, which composes released checkpoints into a single served model that inherits complementary capabilities without additional training (Yang et al., 2026, 2025; Abdin et al., 2024). A standard recipe interpolates between two released checkpoints along a merge path indexed by a coefficient λ (Wortsman et al., 2022; Yadav et al., 2023; Yu et al., 2024), and existing recipes commit to one point on this path for all incoming queries. This commitment is poorly suited to multi-objective serving: distinct objectives attain their optima at distinct points on the merge path, so any single checkpoint very likely compromises at least one of them, and the served decision admits no post-hoc inspection or attribution. As Figure 1 illustrates, a checkpoint that is optimal in aggregate may be suboptimal for an individual query, and queries may prefer different regions of the same merge path.

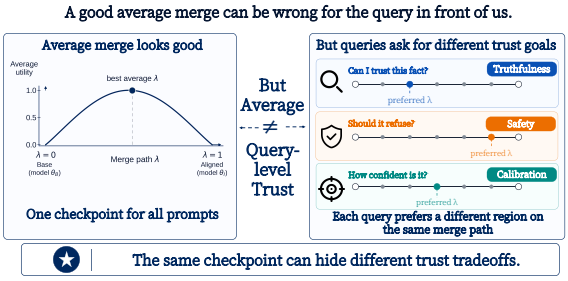

Figure 1: Introduction phenomenon. A checkpoint that is best on average can be the wrong local choice for a query's trust objective; different queries can prefer different regions of the same merge path.

We provide systematic empirical evidence that this heterogeneity is not merely conceptual (Figure 2). On the four checkpoint pairs we study, per-axis utility along the merge path exhibits two coupled patterns: the optima for truthfulness, safety, and calibration (Lin et al., 2022; Hu et al., 2025; Röttger et al., 2024) fall at distinct values of λ on most pairs, and even where they coincide in aggregate, per-query optima remain broadly distributed across many candidate merges (§3.1). We refer to this empirical pattern as *trust-axis anisotropy*: each trust objective favors a different region of the merge path, and the preferred region further varies across queries within a given objective.

Anisotropy reformulates the deployment problem from "which single merge?" to "which merged checkpoint should be served for a given query?". This formulation does not yield to a single end-to-end router, because per-query routing room is thin on some objectives, where naively trained routers overfit, and broad on others, where a single fixed checkpoint leaves utility on the table. PaSM addresses this challenge by treating the per-query merge decision as a measurement problem over a finite bank of candidate merged checkpoints sampled from the merge space. For each query and each objective, PaSM evaluates on held-out data whether a learned specialist outperforms the best fixed checkpoint by a margin sufficient to justify routing; when this condition holds, the specialist is served, and otherwise the validated fixed checkpoint is. Every served decision is logged on a per-query *Trust Card* that records which objectives were routed, which checkpoint was served, and the wall-clock serving cost, rendering per-query routing auditable rather than leaving the served decision opaque to post-hoc inspection.

This paper makes three contributions. **(1) Per-query trust headroom diagnostic** (§3). We provide a per-query, per-axis measurement of routing room on the LLM merge path that, to our knowledge, has not been reported in prior weight-merging work, finding that calibration carries the richest measured signal while truth and safety are more selectively routable, with the routable fraction varying sharply across the four pair types we study. **(2) Per-query serving with validation gating** (§4). We propose PaSM, which builds a Pareto Trust Action Bank, gates routing by validated per-axis headroom, validation-selects specialists per cell, and records each served decision on a per-query Trust Card. **(3) Empirical evaluation with branch accounting** (§5). Across 336 audit cells (4 pairs $\times$ 3 benchmarks $\times$ 28 preferences), PaSM lifts mean utility by $+6.74$ pp over the strongest feasible baseline Mediator and $+2.17$ pp over the best fixed single-λ baseline, certifies 54 paired-bootstrap and 21 max-T cells ($3.4\times$ the strongest single router and $4.2\times$ the strongest external proxy), at 1.06 ms median cached serving latency.

2 Related Work

Weight-space model merging reuses released checkpoints under a new objective. Closest work differs along three dimensions: what it composes, when it conditions on the input, and whether the served action is auditable after the fact.

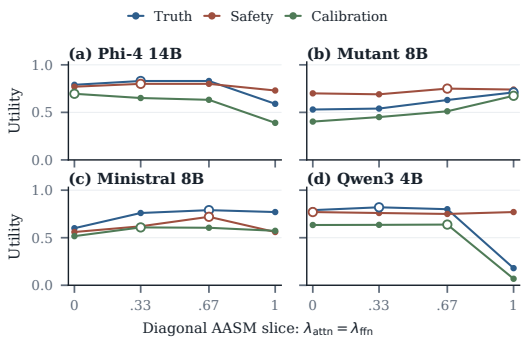

Figure 2: Trust axes do not share one merge optimum. Mean held-out utility on the diagonal trust action bank slice ($\lambda_{\text{attn}} = \lambda_{\text{ffn}}$) for each of the four base-variant pair types; open markers mark the aggregate peak for each axis within the measured slice. Across the four base-variant pair types, no aggregate peak coincides across all three trust axes; so any single fixed merge very likely sacrifices at least one trust axis. This per-pair anisotropy motivates the per-axis Pareto strata of the finite trust action bank constructed in §4.1.

Static and input-conditioned merging. Static recipes such as task arithmetic, TIES, DARE, and Fisher merging establish the weight-space setting (Wortsman et al., 2022; Ilharco et al., 2022; Yadav et al., 2023; Yu et al., 2024; Matena and Raffel, 2022); a recent survey organizes the field by foundations, strategies, and deployment scenarios including safety and alignment (Song and Zheng, 2026). A second wave makes the merge depend on the input: WEMoE (Tang et al., 2024), Twin-Merging (Lu et al., 2024), DaWin (Oh et al., 2025), BD-Merging (Xie and Lyu, 2026), Expert Merging (Zhang et al., 2025), and Mediator (Lai et al., 2025) route or optimize the merge per instance. Pareto Merging (Chen and Kwok, 2024) learns a preference-conditioned continuous Pareto head that synthesizes a merge vector per query. Like other input-conditioned methods, it produces a freshly synthesized weight point on every call, rather than an auditable choice drawn from a measured set of candidate actions on the merge path.

Trustworthiness-oriented merging. A trustworthiness-focused line composes for safety and alignment (Hammoud et al., 2024; Djuhera et al., 2025; Tekin et al., 2026; Ma et al., 2025; Thakkar et al., 2025). Hu et al. (2025) explicitly trace the alignment-calibration trade-off via merging, while multi-objective recipe search (Zhou et al., 2026; Mencattini et al., 2025; Liu et al., 2025) and mergeability analysis (Rahamim et al., 2026) explore broader trustworthiness goals. These works treat trustworthiness as the merging objective, but do not address where in the action

space routing is worth paying for, so per-axis headroom has to be measured on a per-pair basis before any router can be fit.

Calibrated cascading and external-bank routing. Selective prediction and conformal cascading work (Geifman and El-Yaniv, 2017, 2019; Overman and Bayati, 2026; Guo et al., 2026) calibrates router thresholds and defers when uncertain. A parallel line routes among external models or specialist banks (Ong et al., 2024; Wang et al., 2025; Lee et al., 2025; Chuang et al., 2025; Yadav et al., 2024), where the unit of choice is a separately-trained checkpoint rather than an action on a single merge path. PaSM borrows the same calibrated-choice discipline, but its candidates are measured merge actions within a single base-variant pair.

Across these three lines, no prior work simultaneously realizes a finite measured bank with per-axis Pareto strata, per-query routing gated by held-out validation evidence, and a per-query audit record. Static recipes commit to one fixed merge offline. Input-conditioned merging synthesizes a fresh weight point per query without an audit record. External routing achieves per-query audit but routes among separately-trained checkpoints rather than actions on a single merge path. PaSM combines all three on a single merge path; §5 quantifies this against external proxy controls, with extended positioning in Appendix S.

3 Merge Routing to Improve Trust

Before fitting a router, we first measure which queries and axes have routing room above the best-fixed bank action, which we call the *headroom* for routing. The axes are truthfulness on TruthfulQA-MC2, safety on XSTest, and calibration on a per-bank MMLU-Truth confidence composite; the benchmark setup is in §5.1. Throughout, $\mathcal{B}$ denotes the finite action bank and $U_{ax}(q, a) \in [0, 1]$ the per-query utility of action a on axis ax. Two empirical signals matter: axes peak at different merges (§3.1), and per-query routing room above the best-fixed action is sharply axis-dependent across the four pair types we study, with calibration carrying the densest routable signal (§3.2).

3.1 Trust Axes Peak at Different Weights

If one scalar merge coefficient were sufficient for all three trust objectives, their optima should align enough for one global action to be competitive. We observe the opposite. Even on the scalar-λ slice $\theta(\lambda) = (1 - \lambda)\theta_B + \lambda\theta_I$ that varies all weight groups together, aggregate per-axis peak markers move both by pair and by axis (Figure 2), and a per-query top-1 audit confirms that query mass does not collapse onto one globally optimal action.

The serving consequence is concrete. The per-axis Pareto top-1 action diverges across axes within the same pair, and the pattern changes by pair. On Ministral 8B (the running example we trace through §4–§5.5), truth and safety share a top-1 action but calibration is distinct. Mutant 8B is the cleanest case: the three top-1 actions are distinct. On Qwen3-4B and Phi-4 14B, all three axes share a nominal top-1 action, yet cumulative top-1 mass stays moderate. Per-query mass therefore still spreads across other measured actions (Appendix G gives the per-pair identifiers): no single bank action is globally Pareto-optimal across queries on any of the four pairs.

	Truth	Safety	Calibration
Phi-4@14B	13	15	82
Mutant@8B	11	12	31
Ministral@8B	9	12	54
Qwen3@4B	3	7	54

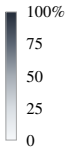

Figure 3: Headroom for routing is axis-asymmetric. Cells show the per-axis share of queries with $h_{ax}(q) > 0.05$ on the measured trust action bank: 31–82% on calibration across the four pairs, 3–13% on truth, and 7–15% on safety. Truth and safety expose stronger routable room on Phi-4 (13% and 15%) than on the other three pairs, which is why PaSM routes those axes whenever validation supports it instead of restricting routing to calibration alone.

3.2 Routing Headroom is Axis-Dependent

Anisotropy alone does not justify routing. A finite action bank earns its routing budget only when the best in-bank action beats the best-fixed bank action by a non-trivial margin on held-out queries. For the finite bank $\mathcal{B}$, we measure this directly via *per-query bank-oracle headroom*: $h_{\text{ax}}(q) = \max_{a \in \mathcal{B}} U_{\text{ax}}(q, a) - U_{\text{ax}}(q, a^{\text{fix}}_{ax})$, where a^{fix}_{ax} is the per-fold train-tuned best-fixed bank action on axis ax. A query is *routable* on that axis when $h_{\text{ax}}(q) > 0.05$, and we denote the resulting per-axis routable-query fraction as $\rho_{\text{ax}} \in [0, 1]$.

Across the four pair types, ρ_{ax} is sharply asymmetric (Figure 3, Appendix F Table 7). Calibration carries the richest measured routing signal: 31–82% of calibration queries are routable across the four pairs, broadly distributed across the per-axis Pareto subset rather than concentrated on one action. Truth and safety expose more selective measured headroom on most pairs: truth

routability ranges from 3–13%, and safety from 7–15%. Phi-4 14B gives the strongest truth and safety readouts, at 13% and 15% routable. Cross-bench checks confirm the asymmetry: BeaverTails and XSTest rank-correlate +0.85 and +0.94 on safety, and HaluEval-QA reproduces the selective truth-routability pattern (Appendix O). The picture is admissibility-gated routing: the system routes wherever pair and axis expose validated room above best-fixed bank actions.

This axis-dependence shapes the routing system in §4. On axes with dense measured headroom, the system serves a validation-selected specialist out of $\mathcal{B}$. On more selective axes, it routes only when validation supports a specialist and otherwise serves the validation-selected fixed bank action.

Headroom upper bound. By definition of $h_{ax}(q)$, for any router $\pi_{ax} : q \to a \in \mathcal{B}$,

$$U_{ax}(q, \pi_{ax}(q)) - U_{ax}(q, a_{ax}^{\text{fix}}) \leq h_{ax}(q). \quad (1)$$

On queries with $h_{ax}(q) \leq \tau$, no router over the same bank improves the reference action by more than τ, which justifies the sparse-axis abstention rule in §4.2. The full bound and bank-expansion monotonicity result are in Appendix D.

3.3 Why Axes Peak Differently

The anisotropy above identifies moving optima, not the weight directions behind them. To probe whether FFN and attention coefficients explain the observed peak shifts, we fit a quadratic regression of bank utility on the per-weight-group λ,

$$\begin{aligned} U_{ax}(\lambda) = \beta_0 + \beta_1 \lambda_{\text{attn}} + \beta_2 \lambda_{\text{ffn}} + \beta_3 \lambda_{\text{attn}}^2 \\ + \beta_4 \lambda_{\text{ffn}}^2 + \beta_5 \lambda_{\text{attn}} \lambda_{\text{ffn}} + \varepsilon, \end{aligned} \quad (2)$$

over 9 pair-axis cells (3 pairs; Appendix H). Two patterns emerge. FFN dominates attention in 8 of 9 cells, with geometric-mean ratio $|\beta_{\text{ffn}}|/|\beta_{\text{attn}}| = 7.6$; the lone exception is Mutant safety, whose surface is near-flat at the XSTest refusal ceiling (Röttger et al., 2024). The interaction coefficient $|\beta_5|$ is large in 5 of 9 cells, so per-axis utility lives in a coupled attention-FFN subspace rather than two independent directions. Both patterns motivate the per-weight-group trust action bank in §4.1, which exposes attention, FFN, embedding, norm, and residual directions separately so per-query controllers can choose among them.

4 Method

The headroom diagnostic from the previous section implies a simple operating rule: route only when validation evidence supports it. We operationalize this rule with the PaSM serving pipeline, which has four components: a finite Pareto Trust Action Bank on the merge path provides the candidate set; a headroom-gated routing rule filters axes by measured per-axis room; per-cell validation-selected specialists, drawn from a candidate pool and frozen before the test fold, score candidates under a preference vector; and a per-query Trust Card records the served decision, the routed axes, and the serving cost. We detail each in turn (Figure 4).

4.1 Pareto Trust Action Bank

The routing system needs a finite bank with per-axis structure. We compile the base-intervention merge path into a Pareto Trust Action Bank with per-axis Pareto strata $\mathcal{B} = \mathcal{B}_{tr} \cup \mathcal{B}_{sf} \cup \mathcal{B}_{cal}$. The parameterization extends scalar-λ to a per-weight-group vector $(\lambda_{\text{attn}}, \lambda_{\text{ffn}}, \lambda_{\text{embed}}, \lambda_{\text{norm}}, \lambda_{\text{other}}) \in [0,1]^5$ with one interpolation coefficient per weight group, letting the bank express FFN-favoring, attention-favoring, and near-endpoint actions as distinct points rather than as one scalar.

Candidate actions come from an 18-entry Latin-hypercube schedule over $[0,1]^5$ (16 LHS samples plus 2 endpoint anchors; Appendix R). Each candidate is materialized as a real merged weight vector $\theta(\boldsymbol{\lambda}) = (\mathbf{1} - \boldsymbol{\lambda}) \odot \theta_B + \boldsymbol{\lambda} \odot \theta_I$ at per-weight-group granularity and evaluated on the training fold per axis. The per-axis bank $\mathcal{B}_{\text{ax}}$ then retains the actions whose per-query utility gain over the default action covers 95% of positive gain mass, with 49–85 axis-labeled rows per pair (per-fold variants and per-axis stratification expand the 18 seed entries; per-pair composition counts in Appendix F). The trust action bank plays two roles: it is the bank over which all routers select, so bank-oracle headroom equals $\max_{a \in \mathcal{B}} U(q, a)$ on it; and it is the bank against which all baselines are adapted, so every router-vs-baseline comparison holds action richness fixed and isolates routing from bank construction.

4.2 Headroom-Gated Routing

Given $\mathcal{B}$, a per-axis *specialist* learns a predicted-utility map $\hat{U}_{\text{ax}}(q, \cdot)$ over $\mathcal{B}$ from pooled hidden-state features at a chosen layer (Appendix A). The bank oracle upper-bounds any induced router by definition, $U_{\text{ax}}^{\text{oracle}}(q) = \max_{a \in \mathcal{B}} U_{\text{ax}}(q, a) \geq U_{\text{ax}}(q, \pi_{\text{ax}}(q))$; we use it as the regret reference for the per-axis specialist in §5.2. A single shared

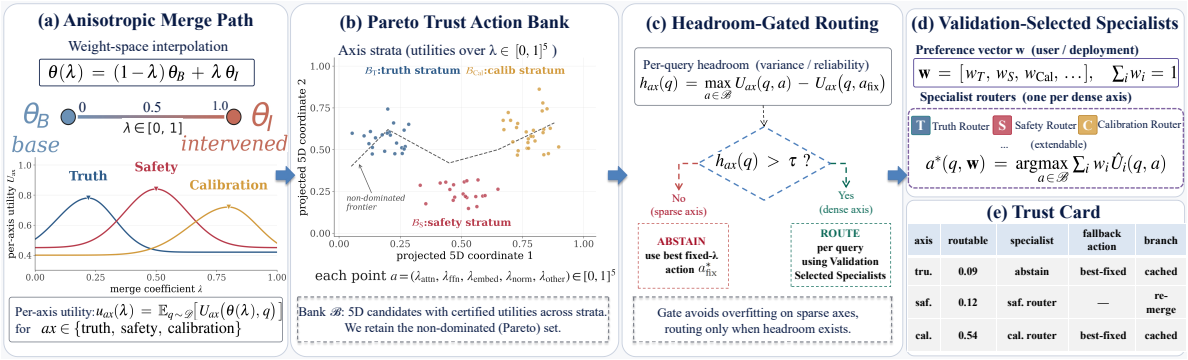

Figure 4: Overview of the PaSM serving pipeline. (a) Anisotropy emerges along the scalar merge path, where different trust axes peak at different values of λ. (b) The Pareto Trust Action Bank expands this path into a frontier over module groups; the illustrative scatter shows the per-axis strata ($\mathcal{B}_T, \mathcal{B}_S, \mathcal{B}_{\text{Cal}}$) and the non-dominated frontier the system retains, not an empirical readout. (c) The headroom gate separates sparse axes, which fall back to the measured-bank best-fixed action, from dense axes that are routed per query. (d) Validation-selected specialists then score candidates under a preference vector $\mathbf{w}$. (e) Each served query is recorded on a Trust Card, supporting per-query audit, branch-aware billing, and selective rollback.

action per preference vector $\mathbf{w} = (w_{tr}, w_{sf}, w_{cal})$ on the 2-simplex (axis weights) is reported as a joint-regret diagnostic in Appendix P.

The headroom-gated routing rule binds routing to measured headroom. It routes when measured headroom supports a per-query specialist (one per axis), and otherwise serves the fixed bank action for that axis. Concretely, on axes with enough measured headroom and validation support, it selects a per-query learned router from a candidate family $\mathcal{F}$ of 24 controllers including TextKNN-Residual, GeoDistill-Sharing, PACR-MLP, BetaHyper, and SAPAR (family-group summary in Appendix B; method-family/ coverage summary in Appendix N). Empirically, calibration meets this condition across all four pairs. Truth and safety on Mutant 8B, Ministral 8B, and Qwen3-4B instead typically serve a^{fix}_{ax}, except Ministral safety's re-merge specialist (§5.3); on Phi-4 14B both axes carry the strongest non-calibration headroom and route through the same validation rule:

$$\pi_{ax}(q, \mathbf{w}) = \begin{cases} \hat{\pi}_{ax}(q, \mathbf{w}), & ax \in \mathcal{D}, \\ a^{\text{fix}}_{ax}, & ax \notin \mathcal{D}, \end{cases} \tag{3}$$

where $\mathcal{D}$ is the set of axes admitted by measured headroom and validation support, and $\hat{\pi}_{ax}(q, \mathbf{w}) = \arg\max_{a\in\mathcal{B}_{ax}} \hat{U}_{ax}(q, a; \mathbf{w})$. For a shared bank action under preference $\mathbf{w}$, the Trust Card composes per-axis predicted utilities:

$$a^*(q, \mathbf{w}) = \operatorname*{argmax}_{a\in\mathcal{B}} \sum_{\text{ax}\in\{tr,sf,cal\}} w_{\text{ax}}\, \hat{U}_{\text{ax}}(q, a), \tag{4}$$

where $\hat{U}_{\text{ax}}(q, a)$ is per-query predicted utility on dense axes and defaults to a per-fold train-mean constant on selective axes where no router is fit.

Algorithm 1: PaSM per-cell selection.

Input: pool $\mathcal{K} = \{\pi^{(k)}\}_{k=1}^{K}$ of $\mathbf{w}$-conditioned specialists; cells $\mathcal{C} = \{(\text{ax}, \mathbf{w})\}$; splits $\{(\text{tr}_f, \text{val}_f, \text{ts}_f)\}$
Output: per-cell selected router Π^{PaSM}

```
foreach fold f, cell (ax, w) do
    foreach π^(k) ∈ K do
        fit on tr_f
        s^(k) ← Ū_ax(val_f; π^(k)(·, w))
    end
    Π^PaSM(f, ax, w) ← π^(arg max_k s^(k))
end
return Π^PaSM on ⋃_f ts_f
```

The headline tables use Equation (3); the shared-action schema is audited in Appendix P. Axis-specificity is therefore forced by the headroom diagnostic rather than imposed by the method: if an axis has measured room and validation support, it can route per query; otherwise the cell retains its validation-selected fixed bank action. The default *calibration head* is fixed before testing by validation cell-win count, giving TextKNN-Residual at $L = 16$ with bandwidth $\eta_{\text{TK}} = 0.3$ on three of four pairs (App. I reports the η_{TK} sweep at $L = 16$).

4.3 Validation-Selected Specialists

The single-head rule leaves headroom on the cells where a different specialist wins. PaSM, the Pareto Specialist Mixture of routers, closes that gap by validation-selecting one specialist per cell from a candidate pool $\mathcal{K}$ (Algorithm 1). The resulting *mixture* is across-cell (one specialist per (axis, $\mathbf{w}$) cell, argmax-selected), rather than a per-query weighted averaging. Per-cell counts use per-axis utility; preference-vector regret is in Appendix P.

Validation rule and pool. The validation score averages non-NaN observed-row utilities over val_f,

with $> 50\%$ validation coverage required and empty cells defaulting to best-fixed. The pool $\mathcal{K}$ expands $\mathcal{F}$ by per-pair layer ($\{8, 16, 24\}$), seed ($\{2, 3, 42\}$), and audited re-merge controllers; each branch's pool is fixed before test-fold evaluation. Candidate predictions are out-of-fold, and PaSM selects specialists on validation folds before scoring held-out folds. Because candidates come from their own 5-fold CV rather than a fully nested selector, we report a fixed validation-selected specialist per cell and include strict nested-CV anchors in Appendix W. Entropy-gated variants are retained only as negative ablations and reported for completeness in Appendix Q (Table 13).

4.4 Trust Card

With one specialist selected per cell on validation evidence, the remaining design question is what gets recorded at serving time so the routing decision can be inspected and re-examined later. PaSM exposes two serving branches that share one validation rule. The cached branch serves predictions from precomputed bank outputs, and the re-merge branch materializes a fresh merged checkpoint on demand for continuous-λ points outside the cached bank. Each served query is then recorded on a per-query Trust Card, the system's audit object. The Card has one row per trust axis and five fields per row, namely the measured routable fraction ρ_{ax} from §3.2, the validation-selected specialist recorded as *abstain* when the axis does not route, the bank's best-fixed action used as fallback, the routing branch taken, and the wall-clock cost. With this record, an operator can inspect why a specialist was preferred over the fixed action for any given query, set per-branch latency contracts that separate cached and re-merge serving, and roll back a single trust axis without disturbing others. §5.5 walks through a worked Ministral example.

4.5 A Guardrail Against Format Collapse

Strict-format judges assign zero utility to malformed outputs, so format collapse silently degrades trust scores while looking like a content failure. The risk is empirical: Figure 5 (§5.2) shows a format cliff on Qwen3-4B past $\lambda \approx 0.67$. The proposition below states a simple condition (call it FC: a malformed output gets zero utility) under which PaSM's selection rule provably cannot pick a collapsed action when the bank itself contains a non-collapsed one.

Proposition (Pointwise retention margin under format–utility coupling)**.** *Fix a pair, fold, axis ax, preference $\mathbf{w}$, and bank $\mathcal{B}$ whose actions have utilities $U_{ax}(q, a; \mathbf{w}) \in [0, 1]$. Assume* format–utility coupling *(FC): there exists a collapse indicator $\kappa(q, a) \in \{0, 1\}$ such that $\kappa(q, a) = 1 \implies U_{ax}(q, a; \mathbf{w}) = 0$. For any frozen score map $\hat{U}$ used by PaSM's specialist, let $\hat{a}(q) = \arg\max_{a \in \mathcal{B}} \hat{U}_{ax}(q, a; \mathbf{w})$ be the selected action on test query q, and suppose the pointwise η-calibration condition $\sup_{a \in \mathcal{B}} \left| \hat{U}_{ax}(q, a; \mathbf{w}) - U_{ax}(q, a; \mathbf{w}) \right| \leq \eta(q)$ holds at this test point q. Then*

$$U_{ax}(q, \hat{a}(q); \mathbf{w}) \geq \max_{a \in \mathcal{B}} U_{ax}(q, a; \mathbf{w}) - 2\eta(q). \tag{5}$$

In particular, if $\max_{a \in \mathcal{B}} U_{ax}(q, a; \mathbf{w}) > 2\eta(q)$, then $\kappa(q, \hat{a}(q)) = 0$ and the selected action retains positive utility.

This bounds the per-query effect of format collapse and addresses one collapse mode, not global capability; complete proof, GSM8K capability sweeps, and selection-λ distributions are in Appendices D, T.

5 Experiments

5.1 Experimental Setup

We instantiate four public base-variant pairs that cover the released checkpoint-family patterns motivating the serving problem: Mutant 8B (Yang et al., 2026), Ministral 8B, Qwen3-4B (Yang et al., 2025), and Phi-4 14B (Abdin et al., 2024) (pair descriptions and benchmark definitions in Appendix E; Table 1 abbreviates the pairs as Phi-4@14B, Mutant@8B, Ministral@8B, and Qwen3@4B). We evaluate the three axes introduced in §3 on TruthfulQA-MC2 (Lin et al., 2022) for truthfulness, XSTest (Röttger et al., 2024) for safety, and a per-bank MMLU-Truth (Hendrycks et al., 2020) confidence composite for calibration; HaluEval-QA (Li et al., 2023) and BeaverTails (Ji et al., 2023) serve as cross-bench diagnostics rather than optimization targets. Each pair contributes 84 cells (3 axes $\times$ 28 preference weights, giving 336 cells total) where every cell aggregates over a preference vector $\mathbf{w} = (w_{tr}, w_{sf}, w_{cal})$ on the 2-simplex with balanced default $\mathbf{w}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and reports a paired-bootstrap CI (Efron, 1992) ($B = 1000$, cluster $= (\text{fold}, q)$) together with a per-pair Westfall–Young max-T correction (Westfall and

	Phi-4@14B			Mutant@8B			Ministral@8B			Qwen3@4B			Summary			
Method	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Avg**	**Gain**	**PB**	**maxT**
best fixed single-λ	79.4	79.1	65.9	70.1	73.3	63.5	80.1	71.3	58.5	79.4	74.8	**62.6**	71.51	0.00	–	–
DaWin	43.5	64.6	60.5	69.4	75.0	41.6	79.2	66.0	59.1	18.7	73.0	12.9	55.29	-16.17	13	8
Mediator	75.5	77.6	68.2	58.9	75.0	44.9	70.6	71.5	48.9	75.2	77.1	59.9	66.94	-4.52	11	5
WEMoE	77.8	75.9	69.4	58.0	74.3	43.7	69.1	68.4	50.0	73.9	77.1	60.1	66.48	-5.02	8	5
BD-Merging	75.1	77.0	67.8	58.5	75.0	45.4	68.9	70.1	49.4	73.3	**77.4**	59.0	66.41	-5.10	8	4
H^3Fusion	75.0	77.0	68.7	60.0	75.0	43.5	71.6	70.0	49.8	73.0	77.0	59.3	66.65	-4.85	7	4
PaSM full	**80.5**	**86.1**	**69.4**	**70.4**	**75.4**	**66.8**	**80.1**	**76.6**	**62.0**	**79.7**	74.8	62.2	**73.68**	**+2.17**	**54**	**21**

Table 1: Main 336-cell audit dashboard. Per-pair Truth/Safety/Calibration utilities (4 pair types $\times$ 3 axes) with the Summary block aggregating, over all 336 cells, average utility, gain over the best fixed single-λ baseline, paired-bootstrap (PB) positives, and Westfall–Young max-T positives. The best fixed single-λ and the five external proxy baselines are defined in §5.1. Bold and underline mark best and second-best numeric values per column.

	Phi-4@14B			Mutant@8B			Ministral@8B			Qwen3@4B			Summary			
Ablation	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Tru.**	**Saf.**	**Cal.**	**Avg**	**Gain**	**PB**	**maxT**
Full PaSM	80.5	**86.1**	**69.4**	**70.4**	**75.4**	66.8	80.1	**76.6**	**62.0**	**79.7**	74.8	**62.2**	**73.68**	**+2.17**	**54**	**21**
w/o re-merge	80.5	86.1	69.4	70.4	75.4	66.8	79.0	69.6	60.5	79.7	74.8	62.2	72.88	+1.37	45	17
w/o per-cell sel.	**85.6**	79.9	64.4	66.2	66.6	57.9	70.2	64.9	56.7	73.5	72.2	53.4	67.64	+0.43	25	12
w/o axis heads	79.9	81.0	66.6	69.4	72.5	**67.4**	**80.3**	70.1	61.0	74.8	**78.0**	57.2	71.51	+0.05	16	10

Table 2: Component ablation. Cumulative removal of PaSM components (re-merge branch, per-cell validation selection, and axis-specific heads) on the same 336-cell grid as Table 1. PB and max-T columns report paired-bootstrap and Westfall–Young family-wise corrected positive-cell counts on the shared bank.

Young, 1993) ($B = 5000$ sign-flip) against the per-fold train-tuned best-fixed bank action a_{ax}^{fix} on the same trust action bank used by every router. Baselines comprise the *best fixed single-λ baseline*, which serves the per-pair validation-tuned best single bank action uniformly across queries and acts as a strong oracle within the universal-single-action class. We further compare against five *feasible external proxy* controls adapted to the same bank: WEMoE (Tang et al., 2024), Mediator (Lai et al., 2025), BD-Merging (Xie and Lyu, 2026), H^3Fusion (Tekin et al., 2026), and DaWin (Oh et al., 2025) (per-method mechanism-matched adaptation details in Appendix M). Internal single router ablations are in Table 2.

5.2 Main Results

Across the 336 cells, PaSM lifts mean utility by $+\mathbf{6.74}$ **pp** over the strongest feasible baseline Mediator and by $+\mathbf{2.17}$ **pp** over the best fixed single-λ baseline (with $+11.53$ pp average gain on paired-bootstrap-positive cells), certifying 54 **paired-bootstrap and** 21 **max-T cells**. The 54 paired-bootstrap cells are $3.4\times$ TextKNN's 16 and $4.2\times$ DaWin's 13; the max-T count is 21 versus 10 and 8. All ratios are matched-cell: every positive cell shares held-out folds, bank, fallback, and correction family. Table 1's pair-axis blocks show where utility moves and the Summary block aggregates over the full grid; served action and branch cost are audited in §5.5. The pre-declared re-merge pool admits Mutant and Ministral on-demand controllers, and gains concentrate on calibration (26 cells) and safety (21 cells), echoing the per-axis headroom asymmetry diagnosed in §3.2 and visualized in Figure 3. We also report GSM8K as a capability-retention scope check rather than as a third headline table, since it measures general task performance rather than the three trust axes optimized above. On Phi-4, PaSM's mean serving $\lambda \approx 0.34$ sits on the rising shoulder of GSM8K where merge accuracy strictly improves over base by $+2$ to $+8$ pp; on Qwen3-4B, the serving $\lambda \approx 0.32$ sits well before the per-pair generation cliff at $\lambda \approx 0.67$ (Figure 5; Appendix T), consistent with the format-collapse guardrail of §4.5. Native LED-Merging in-place runs (Appendix V) cover all four pairs and confirm the Table 1 proxy ranking on each of the four pair types we evaluate.

5.3 Where the Gains Come From

The headline is a portfolio gain rather than a single dominant pair. Ministral drives the re-merge gain: its continuous-λ on-demand controller doubles the max-T contribution from 4 to 8 cells. Mutant keeps the cached branch since its

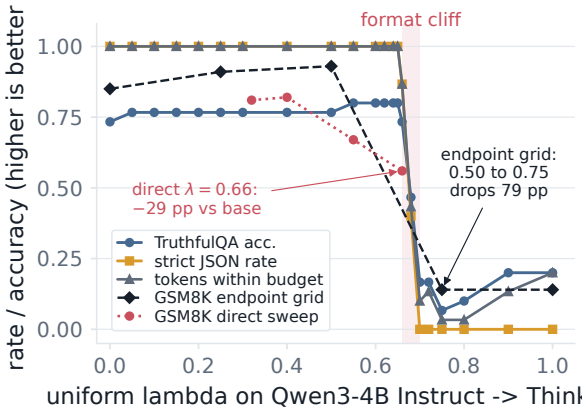

Figure 5: Capability cliff alignment on Qwen3-4B. Four independent measurements on the same merge path (GSM8K-100 accuracy (Cobbe et al., 2021), TruthfulQA accuracy under our strict-JSON evaluator, format-compliance rate, and hit-max-tokens rate) collapse together at $\lambda \approx 0.66$–0.68. The trust-utility cliff used internally by PaSM and the GSM8K capability cliff trace the same transition, which is the empirical content behind the conditional retention statement of Proposition.

re-merge candidate is in the pool but not validation-selected; Phi-4 and Qwen3-4B also serve from the cached branch. At the axis level, re-merge primarily improves safety ($14/4 \rightarrow 21/7$ PB/max-T), calibration remains the densest routable axis ($24/7 \rightarrow 26/8$), and truthfulness stays selective but stable ($7/6 \rightarrow 7/6$). The mix of controller families behind the per-axis lifts is itself non-uniform. TextKNN-style routers dominate calibration cells across all four pairs, while heteroscedastic ensembles and Beta-distribution routers carry most of the safety lift on Ministral and Mutant; truth routing, when admitted, splits between PACR variants and SAPAR rather than concentrating on a single family. Appendix Q decomposes the active validation-selected family per pair-axis block.

5.4 Ablation Study

Table 2 removes PaSM's components cumulatively. Each layer that is taken away costs certified cells and routing precision: 54 PB/21 max-T becomes 45/17 without re-merge, 25/12 without per-cell validation selection, and 16/10 at the pair-shared TextKNN baseline. Certified-cell counts decrease monotonically under cumulative removal, while raw utility need not move monotonically.

Because the headroom gate is tuned for admissibility and cleaner routing, the ablation variants are nested rather than factorial; the table is thus a progressive removal audit. The resulting drop pattern is axis-shaped: removing re-merge mainly costs Ministral's safety lift, removing per-cell validation selection mainly costs the dense calibration block, while the pair-shared TextKNN baseline collapses anywhere routing requires per-axis specialization. Appendix diagnostics cover the multi-seed (Appendix J), re-merge, entropy-gated, and pool-sensitivity checks behind the table.

5.5 Worked Example

Table 3 shows a Ministral 8B Trust Card under balanced preference $\mathbf{w} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Truth abstains to its best-fixed action, safety routes through a re-merge specialist, and calibration through a cached specialist; extended discussion in Appendix U.

axis	ρ_{ax}	specialist $\hat{\pi}_{ax,\mathbf{w}}$	fixed action a_{ax}^{fix}	branch	wall-clock
truth	.09	abstain	best-fixed (truth)	cached	1.06 ms
safety	.12	continuous-λ	—	re-merge	895 ms
calib.	.54	TextKNN-$L16$	best-fixed (calib)	cached	1.06 ms

Table 3: Trust Card worked example. Ministral under balanced $\mathbf{w}$: truth uses the validation-selected fixed action, safety admits a validated re-merge specialist, and calibration routes through a cached validation-selected specialist. ρ_{ax} is the measured routable-query fraction from §3.2.

Serving path	Paid work	Median cost
Selection-only	cached bank/controller lookup	1.06 ms
re-merge cold	fresh re-merge and 32-token forward	2.22 s
re-merge warm	forward after one materialization	895 ms

Table 4: Wall-clock serving cost. Median wall-clock per serving path; the Trust Card separates the low-cost cached route from the materialized re-merge branch. Full p90/p99 latency details remain in Appendix L.

The Trust Card's wall-clock cost field follows the branch: 1.06 ms cached and 2.22 s cold or 895 ms warm for re-merge (Table 4).

6 Conclusion

We introduced PaSM, a trust-aware serving system that turns model merging from a one-shot recipe into a per-query serving pipeline with auditable choice: it compiles a finite Pareto Trust Action Bank on the base-intervention merge path, gates routing by measured per-axis headroom, and records each served decision on a per-query Trust Card. Across four base-variant pair types, PaSM lifts mean utility by $+6.74$ pp over the strongest feasible baseline Mediator, and $+2.17$ pp over the per-pair best fixed single-λ baseline, certifying 54 paired-bootstrap and 21 max-T cells on the 336-cell audit grid at just 1.06 ms median cached-branch latency. Future work should extend PaSM to broader multi-objective merging beyond the three trust axes, to larger model classes, and to streaming Trust Card aggregation for served-action drift detection under shifting query distributions.

Limitations

Scope and cell semantics. The study covers four base-variant pair types with one pair per type. The headline counts therefore describe a fixed audit grid, four pairs, three axes, and 28 preference weights, not the population distribution of all model families or alignment gaps. Within an axis cell, routers share queries, folds, bank, and baselines; across axes, utilities come from different query populations. The Trust Card is therefore a scope-conditioned serving contract, with cross-axis joint regret reported in Appendix P.

Baselines and capability. The external rows in Table 1 are proxy baselines adapted to the same trust action bank. Native LED-Merging and entropy-grid checks now cover all four pairs where feasible and confirm the proxy ranking (Appendix V), but the headline metric remains paired-bootstrap and max-T cell incidence on the 336-cell trust grid. *Trust* here denotes the three measured axes, not alignment as a whole. Capability support is also scoped: the proposition rules out strict-parser format failure only under its calibration conditions, while GSM8K-100 sweeps show per-pair retention surfaces in which PaSM's selected λ values sit away from the measured capability cliffs (Appendix T).

Routing signal, cost, and validation. Calibration carries the richest measured routing signal across the four pair types. Truth and safety have more selective headroom on most pairs and stronger headroom on Phi-4 14B, so the system routes on those axes when validation supports routing and otherwise serves the fixed bank action selected on validation utility. Serving cost is branch-dependent: selection-only lookup is 1.06 ms median, re-merge cold serving is 2.22 s, and warm re-merge forward is 895 ms. PaSM uses a fixed validation-selected specialist per cell from candidate 5-fold predictions; strict nested-CV anchors are reported in Appendix W.

Ethics Statement

This paper studies post-training routing on already-aligned LLMs. Bank actions interpolate between released checkpoints and introduce no new safety training. The bank-oracle headroom diagnostic bounds improvement over λ^*_{fix} only for the measured bank and measured benchmarks; it does not certify production safety on unmeasured behaviors. External safety gates remain necessary before deployment. The datasets used are public research benchmarks; the full reproducibility protocol is in Appendix X.

Use of AI assistants. AI coding/writing assistants were used under author supervision for LaTeX writing polish and paper-wide terminology audit. All claims, experimental design, numerical results, and framing decisions are the authors' own and were verified against source data before inclusion.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Weiyu Chen and James Kwok. 2024. Pareto merging: Multi-objective optimization for preference-aware model merging. *arXiv preprint arXiv:2408.12105*.

Yu-Neng Chuang, Leisheng Yu, Guanchu Wang, Lizhe Zhang, Zirui Liu, Xuanting Cai, Yang Sui, Vladimir Braverman, and Xia Hu. 2025. Confident or seek stronger: Exploring uncertainty-based on-device llm routing from benchmarking to generalization. *arXiv preprint arXiv:2502.04428*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Aladin Djuhera, Swanand Ravindra Kadhe, Farhan Ahmed, Syed Zawad, and Holger Boche. 2025. Safe-merge: Preserving safety alignment in fine-tuned large language models via selective layer-wise model merging. *arXiv preprint arXiv:2503.17239*.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. 2026. Routenlp: Closed-loop llm routing with conformal cascading and distillation co-optimization. *arXiv preprint arXiv:2604.23577*.

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, and Mete Ozay. 2024. Model merging and safety alignment: One bad model spoils the bunch. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13033–13046.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Tiancheng Hu, Benjamin Minixhofer, and Nigel Collier. 2025. Navigating the alignment-calibration trade-off: A pareto-superior frontier via model merging. *arXiv preprint arXiv:2510.17426*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Kunfeng Lai, Zhenheng Tang, Xinglin Pan, Peijie Dong, Xiang Liu, Haolan Chen, Li Shen, Bo Li, and Xiaowen Chu. 2025. Mediator: Memory-efficient llm merging with less parameter conflicts and uncertainty based routing. *arXiv preprint arXiv:2502.04411*.

Seanie Lee, Dong Bok Lee, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. 2025. Saferoute: Adaptive model selection for efficient and accurate safety guardrails in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2053–2069.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. 2025. Sens-merging: Sensitivity-guided parameter balancing for merging large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19243–19255.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935.

Qianli Ma, Dongrui Liu, Qian Chen, Linfeng Zhang, and Jing Shao. 2025. Led-merging: Mitigating safety-utility conflicts in model merging with location-election-disjoint. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21749–21767.

Michael S Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Tommaso Mencattini, Adrian Robert Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. 2025. MergeΘ3: Efficient evolutionary merging on consumer-grade gpus. *arXiv preprint arXiv:2502.10436*.

Changdae Oh, Yixuan Li, Kyungwoo Song, Sangdoo Yun, and Dongyoon Han. 2025. Dawin: Training-free dynamic weight interpolation for robust adaptation. In *International Conference on Learning Representations*, volume 2025, pages 11980–12008.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*.

William Overman and Mohsen Bayati. 2026. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. *Advances in Neural Information Processing Systems*, 38:70305–70337.

Adir Rahamim, Asaf Yehudai, Boaz Carmeli, Leshem Choshen, Yosi Mass, and Yonatan Belinkov. 2026. Will it merge? on the causes of model mergeability. *arXiv preprint arXiv:2601.06672*.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Mingyang Song and Mao Zheng. 2026. Model merging in the era of large language models: Methods, applications, and future directions. *arXiv preprint arXiv:2603.09938*.

Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. 2024. Merging multi-task models via weight-ensembling mixture of experts. *arXiv preprint arXiv:2402.00433*.

Selim Furkan Tekin, Fatih Ilhan, Sihao Hu, Tiansheng Huang, Yichang Xu, Zachary Yahn, and Ling Liu. 2026. H3fusion: Helpful, harmless, honest fusion of aligned llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6993–7013.

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, and Sarath Chandar. 2025. Combining domain and alignment vectors provides better knowledge-safety trade-offs in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277.

Xinyuan Wang, Yanchi Liu, Wei Cheng, Xujiang Zhao, Zhengzhang Chen, Wenchao Yu, Yanjie Fu, and Haifeng Chen. 2025. Mixllm: Dynamic routing in mixed large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10912–10922.

Peter H Westfall and S Stanley Young. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Yuhan Xie and Chen Lyu. 2026. Bd-merging: Bias-aware dynamic model merging with evidence-guided contrastive learning. *arXiv preprint arXiv:2603.03920*.

Prateek Yadav, Colin Raffel, Mohammed Muqeeth, Lucas Caccia, Haokun Liu, Tianlong Chen, Mohit Bansal, Leshem Choshen, and Alessandro Sordoni. 2024. A survey on model moerging: Recycling and routing among specialized experts for collaborative learning. *arXiv preprint arXiv:2408.07057*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chenxu Yang, Qingyi Si, Chong Tian, Xiyu Liu, Dingyu Yao, Chuanyu Qin, Zheng Lin, Weiping Wang, and Jiaqi Wang. 2026. System 1&2 synergy via dynamic model interpolation. *arXiv preprint arXiv:2601.21414*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Dengming Zhang, Xiaowen Ma, Zhenliang Ni, Zhenkai Wu, Han Shu, Xin Jiang, and Xinghao Chen. 2025. Expert merging: Model merging with unsupervised expert alignment and importance-guided layer chunking. *arXiv preprint arXiv:2509.25712*.

Yu Zhou, Xingyu Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. 2026. Hm3: Hierarchical multi-objective model merging for pretrained models. *Advances in Neural Information Processing Systems*, 38:114876–114910.

family group	best pb
2D/5D controller ablations	4
Twin-Merging-style	$0^{\ddagger}$
PACR variants (LR/MLP/MultiL./RankGuard)	7
GeoDistill T-sweep + ablations	13
TextKNN-Residual ($L \times \eta_{\mathrm{TK}}$)	10
BetaHyper / LambdaEBM / SpanPACR	9
SAPAR / EAS-PCR / TS-RFR	8
DaWin-style proxy	13
ContinuousLambda re-merge	17
WEMoE / Mediator / BD-Merging / H^3 proxy	10

Table 5: Method-family summary (best paired-bootstrap cell count across pairs). Controller families listed in App. N. ‡Twin-Merging-style produces a degenerate single-action collapse.

A Hidden Features and Pooling

For each pair, hidden states are extracted at three candidate layers $L \in \{8, 16, 24\}$ (layer 16 is universal best for TextKNN-Residual; layer 8 is best for GeoDistill on Mutant + Qwen3-4B) and mean-pooled across non-padding tokens. Per-axis text features are TFIDF + 256-d SVD; per-axis utility-axis density features for SAPAR / EAS-PCR follow their original formulations.

B What the Controller Pool Contains

The controller-pool ledger groups all evaluated specialist families into method groups: 2D/5D controller ablations, the GeoDistill T-sweep variants, and five proxy/control families adapted to the same bank. Table 5 summarizes the groups; per-row description, hyperparameters, and reproduction commands will be included in the post-acceptance code release.

C PaSM Candidate Pool $\mathcal{K}$

The candidate pool $\mathcal{K}$ for PaSM contains $K = 170$ specialists drawn from $\mathcal{F}$ (Appendix B) by expanding each (method, layer, η_{TK}, seed) tuple where applicable. Pool-sensitivity is reported in Appendix Q.

D Capability Retention by Validation-Utility Argmax

This appendix states and proves the deterministic margin behind the capability-retention check, states the finite-class Hoeffding selection bound used by the stacker-inheritance argument, and records two design observations used in the main text (a headroom upper bound and a bank-expansion monotonicity remark).

Proposition is stated in §4.5. For convenience we restate the pointwise calibration condition assumed there as

$$\sup_{a \in \mathcal{B}} \left| \hat{U}_{ax}(q, a; \mathbf{w}) - U_{ax}(q, a; \mathbf{w}) \right| \leq \eta(q), \quad (6)$$

and we now give the proof.

D.1 Proof of Proposition

The argument is deterministic: it bounds the gap between predicted and true argmax under the pointwise calibration condition (6) and converts it into non-collapse via the format–utility coupling assumption. No probabilistic concentration is invoked at this step.

Fix $(q, ax, \mathbf{w})$ and condition on $\sup_{a \in \mathcal{B}} |\hat{U}_{ax}(q, a; \mathbf{w}) - U_{ax}(q, a; \mathbf{w})| \leq \eta(q)$. Let $\hat{a}(q) = \arg\max_{a \in \mathcal{B}} \hat{U}_{ax}(q, a; \mathbf{w})$ and $a^\star(q) = \arg\max_{a \in \mathcal{B}} U_{ax}(q, a; \mathbf{w})$. Then

$$\begin{aligned} U_{ax}(q, \hat{a}(q); \mathbf{w}) &\geq \hat{U}_{ax}(q, \hat{a}(q); \mathbf{w}) - \eta(q) \\ &\geq \hat{U}_{ax}(q, a^\star(q); \mathbf{w}) - \eta(q) \\ &\geq U_{ax}(q, a^\star(q); \mathbf{w}) - 2\eta(q), \end{aligned}$$

where the first and third lines use the calibration condition and the second uses the definition of $\hat{a}(q)$ as the predicted argmax. This gives (5). If $\max_a U_{ax}(q, a; \mathbf{w}) > 2\eta(q)$, the right-hand side is strictly positive, so $U_{ax}(q, \hat{a}(q); \mathbf{w}) > 0$. The format–utility coupling assumption then forces $\kappa(q, \hat{a}(q)) = 0$, since a collapsed output would imply utility 0. □

D.2 Stacker-inheritance corollary

PaSM validation-selects one specialist per cell: $\hat{k}(\mathrm{ax}, \mathbf{w}) = \arg\max_{k \in \mathcal{K}} \overline{U_{ax}}(\mathrm{val}_f; \pi^{(k)}(\cdot, \mathbf{w}))$, where $\overline{U_{ax}}$ is the validation-fold mean utility. If every $k \in \mathcal{K}$ satisfies the pointwise calibration condition (6) at q with a common bound $\eta(q)$, then so does the selected specialist, and the retention margin (5) applies to the deployed routed action. The common-bound condition is the explicit cost of stacker inheritance: the proposition inherits to PaSM only when the pool $\mathcal{K}$ is uniformly $\eta(q)$-calibrated at q. We do not formally certify this condition; we present it as the standing assumption under which the deterministic margin propagates to the stacked output.

D.3 Finite-class validation-selection bound (Lemma 1)

Pointwise calibration of the deployed specialist is transferred from the pool to the validation-selected

specialist by the common-η assumption in the stacker-inheritance corollary, not by validation-mean selection alone. Separately from that calibration transfer, the validation-mean argmax over the pool is near-best in mean utility by a standard finite-class bound:

Lemma 1 (Finite-class validation selection)**.** *Fix a cell* $(\mathrm{ax}, \mathbf{w})$ *and pool* $\mathcal{K} = \{\pi^{(1)}, \ldots, \pi^{(K)}\}$*. Let* $\mu_k = \mathbb{E}_{q \sim \mathcal{D}_{\mathrm{val}}} U_{ax}(q, \pi^{(k)}(q, \mathbf{w})) \in [0, 1]$ *be the population validation-mean utility of specialist* k *and* $\hat{\mu}_k$ *its empirical estimate over* n *i.i.d. validation queries drawn independently of candidate training. Let* $\hat{k} = \arg\max_k \hat{\mu}_k$ *and* $k^\star = \arg\max_k \mu_k$*. For any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*,*

$$\mu_{\hat{k}} \;\geq\; \mu_{k^\star} - 2\varepsilon_n(\delta, K), \quad \varepsilon_n(\delta, K) = \sqrt{\tfrac{\log(2K/\delta)}{2n}}. \tag{7}$$

The proof is Hoeffding plus a union bound over the K specialists; we omit it. Lemma 1 controls *average* utility of PaSM's selected specialist, not per-query collapse. It is invoked inside the stacker-inheritance argument: if all specialists are pointwise η-calibrated, the validation-selected one is near-best in mean utility by (7), and the pointwise margin then gives non-collapse on the population covered by $\max_a U > 2\eta(q)$.

D.4 Empirical pointwise calibration

The deterministic statement is non-vacuous only on queries where $\eta(q)$ is small enough that $\max_a U_{ax}(q, a) > 2\eta(q)$. We do not certify $\eta(q)$ formally and we are explicit that several specialists in the family-curated pool (in particular the proxy baselines adapted to the same bank) use unclipped predicted scores whose empirical chosen-action error envelopes lie outside the $[0, 1]$ utility range; we report this as a calibration finding rather than as a certificate.

Non-proxy specialists. Restricting to non-proxy router specialists (TextKNN-Residual, GeoDistill, PACR-LR/MLP, BetaHyper, SAPAR, EAS-PCR, TS-RFR, LambdaEBM, SpanPACR, EnsembleAgree; 10–12 per pair), median chosen-action error $|\hat{U}_{ax}(q, \hat{a}) - U_{ax}(q, \hat{a})|$ across the four pairs and three axes is in the range 0.32–0.77; the 95th percentile reaches 1.3–4.0 on the unclipped predicted scores (same envelope caveat as the proxy-baselines paragraph above; computed from per-controller prediction files). Proposition is therefore non-vacuous on cells where the specialist's pointwise error is small and the bank carries some action with utility above $2\eta(q)$; on the calibration axis this condition holds on most queries (calibration utility tends to be high and the routable headroom ρ_{cal} is dense), and the formal guarantee is correspondingly stronger there. On the truth and safety axes, where ceiling-saturated benchmark utilities are binary $\{0, 1\}$ and pointwise specialist error is comparatively large, the deterministic bound is borderline; we therefore rely on the empirical λ-distribution evidence in Appendix T rather than on Proposition alone to support the capability-retention claim on those axes.

Hoeffding reference. For comparison, Lemma 1 at $n = 100$ validation queries per cell, $\delta = 0.05$, and pool size $K = 170$ gives $\varepsilon_n(\delta, K) \approx 0.21$ (corresponding to a mean-utility threshold $\max_a U > 0.42$ per cell, or ≈ 0.54 under an additional 336-cell global union bound). This concerns *average* utility of the validation-selected specialist, not pointwise non-collapse.

Why we keep the proposition. Proposition is the cleanest deterministic statement that captures the design intent of PaSM's selection rule under the format–utility coupling assumption. It does not claim PAC-style certification across the entire pool; the empirical evidence in Appendix T (cliff alignment on Qwen3-4B, selection λ-distribution across all four pairs) is the operational support, and the present submission is honest about the boundary between the proposition and the supporting empirical evidence.

D.5 Cross-fold coupling caveat

PaSM's implemented pipeline uses 5-fold CV candidate predictions, so the validation-fold mean utility $\overline{U_{ax}}(\mathrm{val}_f; \pi^{(k)})$ is computed on a held-out fold whose predictions are produced by a candidate trained on the other four folds; the test fold remains held out, but the candidate that defines the validation mean for fold f may have been trained on a fold that becomes fold f's test queries in another rotation. Lemma 1 requires strict independence between validation-fold queries and candidate-training data; the implemented protocol is therefore a heuristic argument rather than a formal Hoeffding result for the deployed PaSM stacker. We retain Lemma 1 as the conceptual lens behind the selection rule and report only estimates chosen from a fixed bank on validation folds (not strict nested-CV estimates) as the headline claim.

D.6 Why Headroom Controls Routing

Headroom upper bound. For any router $\pi_{ax} : q \to a \in \mathcal{B}$, by the definition $h_{ax}(q) = \max_{a \in \mathcal{B}} U_{ax}(q, a) - U_{ax}(q, a_{ax}^{\text{fix}})$,

$$U_{ax}(q, \pi_{ax}(q)) - U_{ax}(q, a_{ax}^{\text{fix}}) \leq h_{ax}(q).$$

On queries with $h_{ax}(q) \leq \tau$, no router over the same bank can improve the reference action by more than τ. This is the finite-bank justification for the abstention rule in §4.2: on sparse-headroom axes the bank-oracle ceiling is already small, so the measured-bank fixed action is recorded instead of a learned router. We state this as an observation rather than as a result.

Bank-expansion monotonicity (oracle ceiling only). If $\mathcal{B}_0 \subseteq \mathcal{B}_1$, then for every query, axis, and preference vector $\max_{a \in \mathcal{B}_1} U_{ax}(q, a) \geq \max_{a \in \mathcal{B}_0} U_{ax}(q, a)$, so the oracle ceiling is non-decreasing under bank expansion and headroom relative to any reference $a^{\text{fix}} \in \mathcal{B}_0$ is non-decreasing. This monotonicity applies to the oracle/candidate ceiling only; validation-selected held-out cell counts can degrade under pool expansion when the larger pool's val-rule argmax differs from the smaller pool's on some cells. Empirically, the later all-alias pool sensitivity loses one max-T cell relative to the pre-declared $\mathsf{Mut} + \mathsf{Min}$ pool (Appendix W), illustrating this distinction. We therefore admit on-demand re-merge actions under the unchanged validation rule and held-out audit while reporting the later-variant sensitivity as a pool-disclosure, not as a guaranteed non-degrading branch.

D.7 What This Proof Does Not Claim

This appendix proves the capability-retention statement that the paper uses. It does *not* attempt to certify global joint-preference optimality when one prompt must serve all axes; that remains a deliberate scope limit of the headline cell counts. The joint-preference behavior is reported separately as the empirical regret diagnostic in Appendix P.

E Datasets and Models

Pairs. We use four base-variant pairs spanning four intervention *types* (figure/table shorthand labels in parentheses): (i) **Mutant 8B**: Qwen3-8B base paired with a locally-merged Qwen3-8B + reasoning-LoRA variant (LoRA-merge type; both endpoints share Qwen3 architecture); (ii) **Ministral 8B**: Ministral 8B base paired with its released alignment-tuned variant (align-tune type); (iii) **Qwen3-4B**: Qwen3-4B Instruct paired with its Thinking-mode variant (same-family mode-switch); (iv) **Phi-4 14B**: Phi-4 base paired with Phi-4-reasoning (base↔reasoning). Seven checkpoints are publicly released via standard model hubs; the LoRA-merged Mutant variant (i) is obtained by applying a publicly-released LoRA adapter to the public Qwen3-8B base.

Benchmarks. Truth = TruthfulQA-MC2 (Lin et al., 2022); auxiliary HaluEval-QA (Li et al., 2023). Safety = XSTest (Röttger et al., 2024); auxiliary BeaverTails (Ji et al., 2023). Calibration = a per-bank composite over MMLU-Truth (Hendrycks et al., 2020) and alignment baselines, normalized to $[0, 1]$ per axis. The per-query calibration utility is the confidence-aware $1 - |\text{conf} - \text{correct}|$ score, a per-sample calibration measure in the spirit of expected calibration error (Guo et al., 2017) but reported on the $[0, 1]$ utility scale used throughout the paper.

F Trust Action Bank Composition

Per-pair trust action bank composition is reported in Table 6. The per-axis Pareto strata $\mathcal{B}_{tr}, \mathcal{B}_{sf}, \mathcal{B}_{cal}$ are constructed by greedy positive-gain coverage over the 18-entry 5D Latin-hypercube schedule (Appendix R); a single schedule entry can win on multiple axes, so $|\mathcal{B}_{ax}|$ is counted per stratum and may exceed $|\mathcal{B}|$ when the bank is a curated de-duplicated subset. Axis-disjoint share is the fraction of bank actions that win on *exactly one* axis at the population per-axis level (mean 0.81).

Pair	$\lvert\mathcal{B}_{tr}\rvert$	$\lvert\mathcal{B}_{sf}\rvert$	$\lvert\mathcal{B}_{cal}\rvert$	rows	unique
Mutant 8B	28	23	34	85	40
Ministral	19	18	38	75	46
Qwen3-4B	8	12	32	52	39
Phi-4 14B	12	9	28	49	34

Table 6: Per-pair trust action bank composition. *rows* sums per-axis strata (an action winning on multiple axes contributes to each); *unique* counts deduplicated actions. The $|\mathcal{B}|$ used in main-text per-query deployment refers to *rows* (the working bank with axis labels).

Table 7 extends Table 6 with the per-query headroom diagnostic referenced in the main text: routable-query fractions (the share with $h(q) > 0.05$) and mean headroom per axis, alongside the bank-composition columns repeated for direct reading. The §3 heatmap visual (Figure 3) renders the routable-share panel of this table.

Pair	Intervention	Bank		Pareto strata			Routable queries (%)			Mean headroom		
		Rows	Unique	Truth	Safety	Calib.	Truth	Safety	Calib.	Truth	Safety	Calib.
Phi-4@14B	base→reasoning	49	34	12	9	28	**13**	**15**	**82**	.13	.15	.20
Mutant@8B	LoRA merge	85	40	28	23	34	11	12	**31**	.11	.12	.12
Ministral@8B	align tune	75	46	19	18	38	9	12	**54**	.09	.12	.19
Qwen3@4B	same family	52	39	8	12	32	3	7	**54**	.03	.07	.12

Table 7: Bank and headroom diagnostic. Rows report selected-bank size, unique actions, per-axis Pareto-stratum counts, routable-query share ($h > 0.05$), and mean headroom. Bold routable-query fractions mark the axes that drive routing decisions: calibration across all pairs, plus measurable truth/safety room on Phi-4.

G Per-Axis Top-1 Action Distribution

For each pair $\times$ axis, we record *which* bank action is the per-query argmax most frequently across (fold, training query) events (Table 8; 5 folds; per-axis training subset). Top-1 mass is the cumulative count across folds; values > 100 reflect cumulative wins, not a percentage.

Pair (class)	axis → top-1 action	top-1 mass
Mutant	truth → `scalar-02`	110.0
	safety → `5d-attn.5-ffn.5`	58.0
	cal → `5d-attn.5-ffn1.0`	123.8
Ministral	truth → `scalar-13`	80.0
	safety → `scalar-13`	80.0
	cal → `5d-attn.67-ffn.33`	76.3
Qwen3-4B	all three → `5d-attn.5-ffn.5`	22–57
Phi-4	all three → `scalar-13`	4–11

Table 8: Pair-type-conditioned top-1 action concentration. Action codes: `scalar-`k = uniform scalar λ at preset index k in the cached scalar sweep; `5d-attn.X-ffn.Y` = 5D AASM action with $\lambda_{\text{attn}}{=}X$, $\lambda_{\text{ffn}}{=}Y$, and other axes at anchor defaults.

H Per-Cell β-Regression Coefficients

Equation (2) fits a quadratic per-axis utility surface to the fold-aggregated bank-action means. The 8/9 FFN-dominance cells share signs ($\beta_2 < 0$ on truth, $\beta_2 > 0$ on safety; geometric mean $|\beta_2|/|\beta_1| = 7.6$ computed over all nine fits, including the near-flat Mutant-safety cell whose ratio still favors FFN in magnitude). The Mutant safety cell is the lone near-flat fit ($R^2 = 0.37$); we exclude this XSTest-saturated cell from the FFN-dominance claim. The remaining eight cells fit substantially better. Per-cell R^2 values are reported in Table 9.

I 4-Pair Layer $\times$ η_{TK} Sweep (TextKNN-Residual)

Table 10 reports the $L \in \{8, 16, 24\} \times \eta_{\text{TK}} \in \{0.3, 0.5, 0.7, 0.8, 0.9\}$ sweep at $L = 16$ for paired-bootstrap CI cell counts versus λ^*_{fix} on the per-pair 84-cell grid.

J Multi-Seed Robustness

A 4-pair $\times$ 5-layer $\times$ 3-seed grid for GeoDistill T=0.40 yields a best-of-layer 3-seed mean of 34 paired-bootstrap and 10 max-T cells, below the validation-selected PaSM claim ladder but useful as an independent robustness check. Per-pair seed-effect: Mutant + Phi-4 are seed-sensitive (σ across 3 seeds ≥ 2 cells); Ministral + Qwen3-4B are seed-stable.

K Re-merge Continuous-λ Diagnostic

Re-merge performs actual per-query checkpoint re-merge at FFN-predicted unique $(\lambda_{\text{attn}}, \lambda_{\text{ffn}})$ values for each query and re-evaluates downstream utility. This is the only method in our method list that performs real per-query re-merge rather than cached-prediction interpolation or controller-level stacking. Re-merge diagnostics differ by pair: on Phi-4, the single-method pilot reaches 17/84 paired-bootstrap CI cells and 3/84 max-T cells but does not lift the stacked four-pair envelope; on Ministral, the same family of continuous- λ actions lifts the pair from 9/84 pb and 4/84 max-T under selection-only PaSM to 18/84 pb and 8/84 max-T, which raises the four-pair descriptive total from 45 paired-bootstrap and 17 per-pair max-T cells (out of 336) to 54 and 21 cells. We therefore treat the re-merge branch as high-routability with a serving-cost caveat, not as the default low-cost path.

L Re-merge Serving Latency Micro-benchmark

We quantify the branch cost surfaced in the Trust Card (§5.5) by measuring per-query wall time on Qwen3-4B (Instruct $\leftrightarrow$ Thinking), bfloat16 on a single local GPU with background processes co-running. We separate three serving conditions: *Selection-only* (cached bank-prediction lookup, no model inference); *Re-merge cold* (worst-case: each query gets a different routed $(\lambda_{\text{attn}}, \lambda_{\text{ffn}})$ and triggers a fresh in-place re-merge + forward pass); *Re-merge warm* (steady-state: one re-merge, then

Pair	Truth	Safety	Calibration	Readout
Mutant@8B	**1.94**/.89	1.32/.37†	**2.42**/.96	LoRA pair; safety is the only low-fit exception
Ministral@8B	**33.23**/.71	**13.58**/.89	**23.66**/.85	strong FFN-dominance on all axes
Qwen3@4B	**4.88**/.93	**15.48**/.68	**17.18**/.91	same-family pair; all axes FFN-dominant

Table 9: Quadratic mechanism diagnostic. Each cell reports FFN/attention dominance r and fit R^2 as r/R^2 for one pair–axis fit of Eq. 2. Bold marks $r \geq 1.5$; † marks the lone low-fit cell. Dominance ratios and R^2 are summarized in Appendix H.

Pair	.3	.5	.7	.8	.9	shape
Mutant	3	3	1	0	0	low-η_{TK}
Ministral	10	8	5	1	1	decay
Qwen3-4B	1	2	8	8	8	inverted
Phi-4	10	9	8	7	7	decay+floor

Table 10: η_{TK} sweep at $L = 16$; paired-bootstrap CI counts.

multiple queries reuse the materialized checkpoint + forward only).

Path	Median	p90	p99	Throughput
Selection-only lookup	1.06 ms	1.13 ms	1.47 ms	—
Re-merge cold (per-query re-merge; worst-case):				
re-merge step	1317 ms	1387 ms	1475 ms	—
forward (32 tok)	890 ms	962 ms	991 ms	—
total	2220 ms	2339 ms	2427 ms	—
Re-merge warm (one re-merge, subsequent forward-only):				
forward-only steady	895 ms	946 ms	995 ms	**35.7** tok/s

Table 11: Per-query wall time for two PaSM serving paths on Qwen3-4B, bfloat16 on a single local GPU with background processes co-running. $N = 100$ measurements for selection-only and re-merge cold; $N = 58$ for re-merge warm steady-state after a one-shot re-merge. *Throughput* is computed from the warm forward latency on a 32-token generation; the figure below shows how the one-time re-merge cost amortizes once a route is reused.

Figure 6 visualizes the amortization curve. The measurement quantifies the deployment trade-off encoded by c_{branch}: selection-only is millisecond-scale serving on top of cached bank predictions; Re-merge materializes a new merged checkpoint and pays a one-time ≈ 1.3 s re-merge cost. *When multiple queries share the same routing decision* (warm steady-state), forward latency settles at $\approx$ 895 ms / $\sim$ 35.7 tokens per second. This is comparable to running the base or intervention checkpoint alone; the re-merge is amortized across the batch and disappears from per-query overhead. Re-merge is therefore an audit-time worst-case but a deployment-feasible warm-state when routing decisions batch. Pairs at 8B (Ministral) and 14B (Phi-4) scale roughly with parameter count; local background-process noise inflates the cold-state tail (p99/median $\approx$ 1.05–1.10 across conditions), and the order-of-magnitude branch-cost ratios are robust to host noise.

M Mechanism-Matched Proxy Controls

For each prior method (WEMoE, Mediator, BD-Merging, H^{3}Fusion, DaWin) we adapt its core dynamic-gating mechanism to our cached trust action bank evaluator. Implementation summary: WEMoE-style = MoE gate over bank actions trained on hidden features; Twin-Merging-style = shared/exclusive split (*degenerate*; single-action collapse on this bank); DaWin-style = cached-utility entropy gate; Mediator-style = preference-conditioned bank action; BD-Merging-style = λ-decomposition gate at population level; H^{3}Fusion-style = 3-axis MoE over bank actions, no per-query gating. Each proxy uses identical 5-fold splits, hidden-feature pool, and trust action bank; cell counts in Table 1 are comparable only as bank-action diagnostic controls under our evaluator. They are not claims about the original papers' published checkpoint-level performance.

N Controller/HP Grid Coverage

The 114 evaluated controller/HP variants span Mutant 30 + Ministral 35 + Qwen3-4B 26 + Phi-4 23. Methods covered: 2D/5D controller variants, Twin-Merging, PACR (LR/MLP/MultiLayer/RankGuard), the GeoDistill T-sweep with ablations, TextKNN-Residual at $L \in \{8, 16, 24\} \times \eta_{\mathrm{TK}} \in \{0.3, \ldots, 0.9\}$, BetaHyper, LambdaEBM, SpanPACR, SAPAR, EAS-PCR, TS-RFR, a DaWin-style proxy, a continuous-λ re-merge controller, and four proxy baselines adapted to the same bank. The grid also includes validation-selected PaSM claim-ladder rows and the exploratory rows used for the negative ablation in Table 13.

O Cross-Bench Validity Detail

Safety axis: BeaverTails vs. XSTest. On the 5 common bank actions instrumented under both safety benches, Pearson rank-correlation is $+0.85$ (Mutant) and $+0.94$ (Ministral). BT exposes $+9.5$ pp / $+15.5$ pp more routable safety queries than XSTest (21.5% vs. 12%; 27.5% vs. 12%). XSTest saturation is bank-size-bounded, not a structural property.

Truth axis: HaluEval-QA vs. TruthfulQA-MC2 (4 pairs). A 60-query $\times$ top-12 actions diagnostic on every pair shows HaluEval frac_headroom

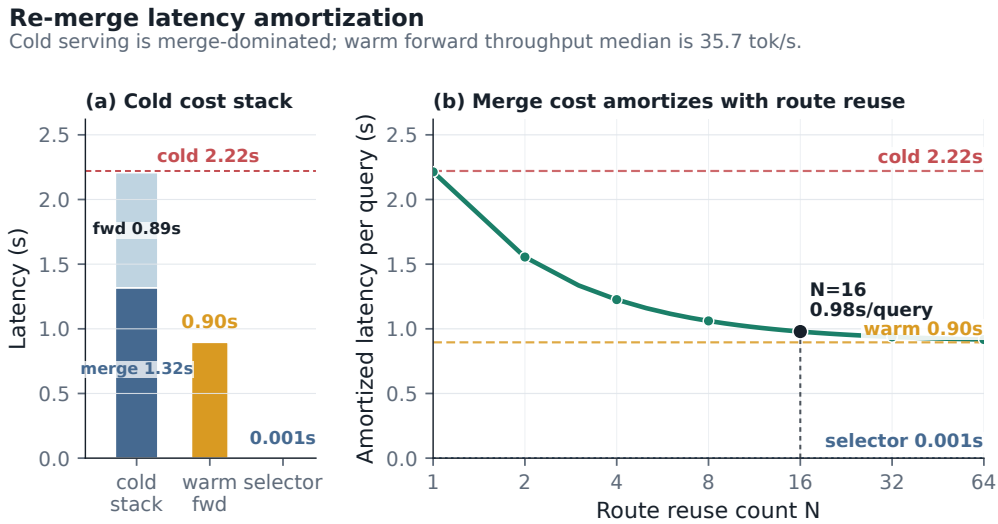

Figure 6: Re-merge serving latency amortizes after route reuse on Qwen3-4B (single local GPU, bfloat16). Left: cold serving is dominated by the one-time merge plus forward pass, with warm-forward and selector references shown on the same seconds scale. Right: if N queries share a re-merge route, the effective per-query latency is $t(N) = t_{\text{select}} + t_{\text{warm}} + t_{\text{merge}}/N$. The marked $N = 16$ point is already near the warm-forward regime, showing that route materialization becomes small once a specialist route is reused across a modest batch.

1.7% / 3.3% / 0% / 1.7% (Mut / Min / Qwen3-4B / Phi-4) versus TQA 11% / 9% / 3% / 13%. HaluEval is more saturated than TQA on every pair tested (one-sided binomial $p \in [0.98, 1.00]$); the truth-axis low-routability finding is *not* a TQA-specific artifact.

Headroom threshold sensitivity. On Mutant + Ministral pairs (truth + safety axes), the per-query bank-oracle headroom H has a near-binary distribution: $\text{frac}(H > .05) = \text{frac}(H > .20) = \text{frac}(H > .50)$ on each axis (Mutant truth $0.11/0.11/0.11$, safety $0.12/0.12/0.12$; Ministral truth $0.09/0.09/0.09$, safety $0.12/0.12/0.12$). The threshold $h(q) > 0.05$ used in the main text is therefore not a knife-edge: the saturated truth+safety axes have either $H = 0$ or $H = 1$ on each query, so any threshold in $(0, 1)$ yields the same routable fraction. On the calibration axis, the distribution is continuous: Mutant cal $0.31/0.14/0.10$ (at $H > .05/.20/.50$); Ministral cal $0.54/0.21/0.15$. The headroom asymmetry conclusion is robust to the 0.05 threshold choice on the saturated axes (§3.2).

P Cost and Joint-Regret Sensitivity

Bank construction cost. $\sim$60–120 GPU-hours per pair (one-time; 18-entry Latin-hypercube schedule, greedy positive-gain bank derivation, and per-bench evaluation).

Per-query inference cost. The selection-only branch adds a router forward pass on hidden features plus a cached $\arg\max$ over $|\mathcal{B}| \leq 85$ candidate actions; the re-merge branch additionally requires checkpoint materialization or an equivalent delta-composition serving path. Appendix L (Table 11) reports the latency micro-benchmark; the main paper therefore treats re-merge as a higher-routability, higher-cost branch rather than a cost-light default.

Joint preference-vector regret sensitivity. A diagnostic sweep over $|\Delta^3| = 28$ preference weights $\times$ 4 pairs $=$ 112 cells, computed as $\text{regret}(\text{pair}, \mathbf{w}) = -\sum_{\text{ax}} w_{\text{ax}} \overline{\text{gain}}_{\text{ax}}^{\text{best-fix}}$ on *both* Trust-Card rules: **(i) selection-only PaSM stacked rule** (val-rule cell-stacked over validation-selected specialists per (axis, $\mathbf{w}$)): negative joint regret on $\mathbf{55/112}$ cells, mean -0.0108; per-pair Mut 15/28, Min 8/28, Qwen3-4B 11/28, Phi-4 21/28. **(ii) headroom-gated single-head rule** (truth+safety carry the measured-bank deterministic action on the three LoRA/align/same-family pairs; TextKNN-Residual $L = 16$ on calibration with η_{TK} chosen by pair type; all three axes route on Phi-4): negative joint regret on $52/112$ cells, mean -0.0019; per-pair Mut 12/28, Min 16/28, Qwen3-4B 0/28, Phi-4 24/28. The stacked rule produces stronger *average* joint regret (-0.0108 vs -0.0019) and dominates on Qwen3-4B (where the single-head headroom gate is weight-flat at 0/28 because all preference vectors there favor the calibration head); the headroom gate is slightly

Pair axis	Gain pp	PB/ maxT	Dom. family	Top/ top-5	Pair PB/maxT
Phi-T	$+1.2$	$3/3$	Ens.	$11.2/25.1$	$15/5$
Phi-S	$+7.0$	$9/1$	Beta	$15.3/26.7$	
Phi-C	$+3.5$	$3/1$	LR	$11.3/23.8$	
Mut-T	$+0.3$	$1/1$	DaWin	$8.8/25.2$	$17/5$
Mut-S	$+2.1$	$3/1$	BD	$18.3/28.0$	
Mut-C	$+3.3$	$13/3$	EAS	$9.7/28.0$	
Min-T	-0.1	$1/1$	re-merge	$8.8/25.3$	$18/8$
Min-S	$+5.3$	$8/4$	re-merge	$24.0/28.0$	
Min-C	$+3.5$	$9/3$	re-merge	$18.2/28.0$	
Qwen3-T	$+0.3$	$2/1$	Ens.	$10.8/25.1$	$4/3$
Qwen3-S	$+0.0$	$1/1$	BD	$8.7/24.5$	
Qwen3-C	-0.4	$1/1$	KNN	$16.0/27.7$	

Table 12: Where the PaSM gains come from. Rows summarize the 28 preference-weight cells for one pair–axis block. T/S/C denote truth, safety, and calibration. Gain is mean utility lift in percentage points; Top/top-5 is selected-cell mass assigned to the dominant family and top five families. Axis totals: Truth $7/6$, Safety $21/7$, Calibration $26/8$ PB/max-T.

stronger on Ministral and Phi-4 where its single calibration head is already preference-aligned. Joint dominance is not claimed.

Q PaSM Sensitivity to Variant Choice

Selection-only branch. The validation-selected selection-only PaSM branch reaches 45 paired-bootstrap and 17 per-pair max-T cells (out of 336) without per-query checkpoint re-materialization.

Adding On-Demand Re-Merge Actions. Adding the Ministral continuous-λ on-demand re-merge controller raises the descriptive four-pair total to 54 paired-bootstrap and 21 per-pair max-T cells (out of 336).

Negative-ablation entropy-gated variants. Exploratory top-2 and entropy-gated variants are retained only as negative-ablation rows after re-evaluation under the reported rule; they are not used as headline performance claims because the corresponding selection rules are not the reported validation rule.

Per-pair, per-axis gain decomposition. Table 12 reports the per-(pair, axis) gain, paired-bootstrap / max-T count, and dominant validation-selected specialist family for the 336-cell audit; axis totals are $7/6$ (truth), $21/7$ (safety), and $26/8$ (calibration) paired-bootstrap / max-T. Although the per-(pair, axis) mean gain in Table 12 can sit near zero on a small number of blocks (e.g., Min-T, Qwen3-C), cell-level PB/max-T positives can co-exist with them because each (pair, axis, preference-weight) cell is tested independently under its own bootstrap distribution; the count is a robustness diagnostic, not a mean-gain alignment.

Configuration	pb / 336	m-T / 336
Current validation-selected claim ladder		
Paper-cached fallback	35	14
Selection-only PaSM (primary)	**45**	**17**
PaSM + Ministral re-merge	54	21
Exploratory entropy-gated variants (negative ablation)		
Entropy gate under reported rule	17	8
Family + entropy under reported rule	14	6
Family + entropy ($q = 0.35$) under reported rule	13	5

Table 13: PaSM claim ladder and negative ablations. Cell counts on the 336-cell grid versus the best fixed single-λ baseline (per-pair validation-tuned single bank action; see Table 1 caption for the Oracle framing). The top block contains validation-selected claim candidates: the paper-cached fallback, the primary lower-cost selection-only branch, and the optional higher-cost Ministral re-merge augmentation. The bottom block records entropy-gated exploratory variants after re-evaluation under the reported rule; they are negative ablations, not performance claims.

R Bank Construction Schedule

The 18-entry 5D Latin-hypercube schedule $\{(\lambda^{(i)}_{\text{attn}}, \lambda^{(i)}_{\text{ffn}}, \lambda^{(i)}_{\text{embed}}, \lambda^{(i)}_{\text{norm}}, \lambda^{(i)}_{\text{other}})\}_{i=1}^{18}$ samples the 5-axis space via Latin hypercube (16 samples) plus base and intervention endpoint anchors (2 entries), totalling 18 entries. This is a budgeted finite serving bank chosen for the per-pair $\sim$ 60–120 GPU-hour bank-construction cost in Appendix P, not a coverage estimate of the continuous 5D space; downstream per-axis Pareto stratification and per-fold variants expand this seed schedule into the 49–85 axis-labeled bank rows reported in Table 6.

S Positioning Detail Across the Mixture of Routers vs. Mixture of Weights Distinction

Table 14 summarizes how each weight-merging neighbor differs from PaSM on composition object and gating signal.

T Capability Retention Evidence

A natural reviewer concern is whether the trust-axis gains reported in Table 1 are obtained by collapsing the base model's general capability. We address this through a deterministic conditional argument (Proposition; Appendix D) together with three pieces of empirical evidence: (a) our trust benchmarks impose format and prompt-following requirements that yield zero trust utility under output collapse; (b) the cached merge landscape on the Qwen3-4B pair exhibits a sharp capability cliff whose location aligns with the trust-utility cliff; (c) PaSM's per-query selections cluster on the safe side of that landscape on all four pairs. We do not claim that capability retention follows from the selection rule alone; the formal guarantee is conditional on pointwise specialist calibration (Proposition), and we are explicit about that boundary.

Trust benchmarks couple to format compliance. Truth, safety, and calibration utility in this paper are measured on prompt-following benchmarks (TruthfulQA, XSTest with BeaverTails diagnostics, MMLU-Truth): each query requires the merged model to follow a JSON-formatted instruction, parse the question, produce a structured answer, and meet a format constraint. A merge action whose output is format-collapsed (outputs fail to parse, hit the max-token cap, or skip the answer field) therefore yields trust utility close to zero. This is the format–utility coupling assumption underlying Proposition. The selection rule then prefers specialists whose actions have high validation-fold trust utility, and under the pointwise calibration condition this preference excludes capability-collapsed actions on covered queries; we do not claim exclusion as an unconditional structural property. Calibration utility is in particular anchored on MMLU-Truth (Hendrycks et al., 2020), the canonical LLM capability benchmark, so the calibration axis is intrinsically a capability-sensitive measurement.

Empirical cliff alignment on the Qwen3-4B pair. Figure 5 (visual) and Table 15 (numbers) report the cached lambda-grid landscape for the Qwen3-4B Instruct $\rightarrow$ Thinking pair on three independent measurements: GSM8K-100 accuracy (Cobbe et al., 2021), TruthfulQA accuracy under our strict-JSON evaluator, and the format-compliance rate on the same evaluator. All three signals collapse together at $\lambda \approx 0.66$–0.68, producing a sharp capability cliff. The trust-utility cliff used internally by PaSM on this pair therefore aligns with the GSM8K capability cliff: empirically, the validation argmax that maximizes trust utility prefers λ-bank regions away from the cliff on this pair. We report this as landscape evidence on one of the four paper pairs, not as a cross-pair guarantee.

Phi-4 no-cliff check. The Qwen3-4B cliff at $\lambda \approx 0.67$ is not universal across the four pair classes. We sweep GSM8K-100 accuracy at $\lambda \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$ on the Phi-4 base$\rightarrow$Phi-4-reasoning pair under the same qwen-compact prompt format. Accuracy is *non-monotone with peak at* λ=0.80: 0.63 (base) $\rightarrow$ 0.65 / 0.71 / 0.73 / 0.79 / **0.83** (peak) $\rightarrow$ 0.71 (λ=1, full reasoning); see Table 16. The drop from λ=0.80 to λ=1.0 is -12 pp, an order of magnitude smaller than the Qwen3-4B cliff (-79 pp from 0.93 to 0.14), and the lowest sweep point (λ=1.0, acc. 0.71) still strictly

Method	Composition object	Distinguisher vs. PaSM
WEMoE (Tang et al., 2024)	MLP weight-ensembling MoE (shared weights/task vectors)	gates weights, not routers
Twin-Merging (Lu et al., 2024)	shared/exclusive merge weights	weight-level decomposition
DaWin (Oh et al., 2025)	entropy-gated mixture coefficients	gating signal = pred entropy, not val-rule selected
BD-Merging (Xie and Lyu, 2026)	per-sample task/layer weight router	weight-level routing, not router-stacking over headroom-gated bank
H^3Fusion (Tekin et al., 2026)	3-axis Helpful-Honest-Harmless (HHH) expert MoE	weight MoE, no abstention diagnostic
Mediator (Lai et al., 2025)	layer-conflict handling + uncertainty-based expert routing	weight-level routing, no Pareto-bank specialization
$MERGE^3$ (Mencattini et al., 2025)	IRT-assisted evolutionary merging-coefficient search	no per-query routing
Headroom gate (ours)	per-axis specialist routers	Mixture of Routers; structured abstention
PaSM (ours)	validation-selected router-branch stacking	router-level stacking gated by per-axis bank-oracle headroom

Table 14: Compact positioning. All Mixture of Weights neighbors compose *checkpoints/weights* with per-input gating; closest semantic overlaps to PaSM are DaWin (per-sample coefficient selection), Mediator / BD-Merging (uncertainty / task-layer routing), and LED-Merging / SafeMERGE (safety-utility composition). PaSM's specific delta is the combination of (a) per-axis bank-oracle *headroom* diagnostic computed before routers are fit, (b) axis-level *abstention* on sparse-headroom axes, and (c) *validation-selected stacking* of axis specialists per $(\text{axis}, \mathbf{w})$ cell over a measured Pareto bank. The five proxy controls in Table 1 adapt each method's core dynamic-gating mechanism to our cached trust action bank evaluator; they are diagnostic adaptations, not faithful checkpoint-level reproductions of the originals on their native tasks.

λ	0.50	0.65	0.66	0.68	0.70	1.00
GSM8K-100 acc.[‡]	0.93	0.93	–	0.14	0.14	0.14
TruthfulQA acc.	0.77	0.80	0.73	0.47	0.17	0.20
format-compliance rate	1.00	1.00	0.87	0.40	0.20	0.20
hit-max-tokens rate	0.00	0.00	0.13	0.57	0.90	0.80

Table 15: Capability-cliff alignment on the Qwen3-4B Instruct → Thinking pair under uniform-λ merging. GSM8K, TruthfulQA, and format-compliance collapse together at $\lambda \in [0.66, 0.68]$. The trust-utility cliff used by PaSM therefore aligns with the capability cliff: maximizing trust utility on this pair empirically prefers λ-bank regions away from the cliff. [‡]GSM8K is reported on the cached $\{0, 0.25, 0.5, 0.75, 1\}$ grid (no point at λ=0.66); λ=0.75 is below the table column but matches λ=0.70 (acc. 0.14).

λ	0.00	0.20	0.40	0.50	0.60	0.80	1.00
GSM8K-100 acc.	0.63	0.65	0.71	0.73	0.79	**0.83**	0.71

Table 16: Phi-4 base → Phi-4-reasoning GSM8K-100 accuracy under uniform-λ merging (qwen-compact prompt, seed 42). Non-monotone with peak at λ=0.80. The drop from peak to full-reasoning (λ=1) is -12 pp, an order of magnitude smaller than the Qwen3-4B cliff in Table 15; the worst measured λ on Phi-4 still exceeds the base by $+8$ pp.

exceeds the base (λ=0.0, acc. 0.63). There is no *capability-collapse* region anywhere on the Phi-4 sweep; reasoning interpolation strictly improves math reasoning over base on every measured λ. PaSM's mean per-query λ on Phi-4 is ≈ 0.34 (selection distribution below), sitting in the rising-shoulder of the curve (≈ 0.65–0.71 accuracy), above the base by $+2$ to $+8$ pp. This contrast with Qwen3-4B reinforces the §3.1 framing that the capability behavior along the merge path *depends on pair type*.

Endpoint-specialist routing across pairs. Table 17 reports the endpoint anchor utilities used to interpret the selector. The dominant endpoint is pair- and axis-conditioned: Qwen3-4B and Phi-4 are base-dominant on truth and calibration, Mutant is intervention-dominant on all three trust axes, and Ministral is intervention-dominant on truth with smaller safety and calibration gaps. Table 18 then audits the validation-selected action distribution. The "mid-λ" bucket is strict: all five AASM axes must fall in the joint middle range; preset actions can still be mid on attention/FFN while remaining endpoint-like on the other axes. The resulting 0.0% strict-joint mid-λ mass across all four pairs is therefore selector sparsity in that strict bucket, consistent with functional-axis winners clustering at specialist extremes.

PaSM-selected actions stay below the cliff. Table 19 reports the per-pair distribution of PaSM's per-query action selections, expressed via

Pair	Truth (B / I)	Safety (B / I)	Cal (B / I)
Qwen3-4B	**0.79** / 0.16	0.78 / 0.78	**0.64** / 0.12
Mutant 8B	0.53 / **0.71**	0.70 / **0.74**	0.44 / **0.71**
Ministral 8B	0.60 / **0.79**	0.51 / 0.53	0.51 / **0.58**
Phi-4	**0.79** / 0.29	0.78 / **0.82**	**0.70** / 0.25

Table 17: Per-pair endpoint utility for anchor-B (base, λ=0) and anchor-I (intervention, λ=1) on truth, safety, and calibration axes ($N = 100$ per endpoint cell). Bold marks the larger endpoint value when the difference is nonzero.

Pair	Low-λ	Mid-λ	Cliff	B-anchor	I-anchor	Preset
Qwen3-4B	55.3%	**0.0%**	6.6%	0.0%	7.9%	30.2%
Mutant 8B	65.5%	**0.0%**	3.5%	0.0%	17.2%	13.7%
Ministral 8B	51.8%	**0.0%**	7.7%	0.0%	0.0%	40.4%
Phi-4	44.2%	**0.0%**	9.4%	5.8%	8.3%	32.4%

Table 18: Pre-declared val-rule selector-mass audit by chosen-action λ bucket, aggregated over the family-curated controller pool (126–143k action choices per pair; numerically unchanged from the selection-only audit because the Mutant re-merge candidate is selected on no cells). "Mid-λ" uses the strict five-axis joint definition; base and intervention anchors are exact all-axis endpoint matches, and preset contains named continuous-λ or AASM preset actions outside the strict buckets.

the weighted mean of the action-level attention and FFN merge coefficients. For Ministral, Qwen3-4B, and Phi-4, the mean selected λ values stay between 0.31 and 0.47, well below the ~ 0.67 uniform-λ cliff observed on Qwen3-4B. The Mutant pair has a higher mean λ, but its selections concentrate at the intervention anchor (17.2% of choices) and the symmetric high-λ bank cells; this is the regime where the Mutant-Reasoning intervention itself defines the capability behavior, not a collapse zone. Read together with Table 18, these means separate two notions: some selected presets are mid on attention/FFN, but strict joint-mid selection across all five AASM axes remains absent. Thus the selector evidence supports endpoint/preset specialists rather than a generalist joint-mid policy.

Why Interpolation Can Preserve Capability. Beyond our own evidence, weight-space interpolation between same-family checkpoints is a standard post-training reuse mechanism with documented capability-preserving regimes. Model soups (Wortsman et al., 2022) preserve or improve accuracy under uniform averaging; task arithmetic (Ilharco et al., 2022) composes task vectors without retraining; TIES-Merging (Yadav et al., 2023) recovers per-task accuracy via sign-conflict resolution; sensitivity-based merging (Liu et al., 2025) preserves task capability via importance-weighted merging; SafeMERGE (Djuhera et al., 2025) and LED-Merging (Ma et al., 2025) target

pair	$\overline{\lambda}_{\text{attn}}$	$\overline{\lambda}_{\text{ffn}}$	% B-anchor	% I-anchor	% preset
Mutant	0.57	0.71	0.0	17.2	13.7
Ministral	0.34	0.47	0.0	0.0	34.6
Qwen3	0.34	0.31	0.0	7.9	30.2
Phi-4	0.34	0.34	5.8	8.3	32.4

Table 19: PaSM per-query selection statistics. $\overline{\lambda}_*$ are weighted mean attention/FFN coefficients across all selected actions ($n = 126\text{k} - 143\text{k}$ per pair). The B-anchor / I-anchor columns report the share of selections that fall exactly on the base or intervention endpoint; the preset column reports the share on named bank presets. This mean-coefficient view complements the stricter bucket audit in Table 18: presets may be mid on attention/FFN while not satisfying the five-axis joint mid-λ definition.

the safety–utility composition problem and demonstrate that capability can be retained under safety-oriented merging. PaSM is consistent with this line of work: it draws per-query actions from a validation-curated bank that survives the same family of selection guarantees rather than from arbitrary interpolation points.

Direct evidence at PaSM's operating point. On Qwen3-4B we additionally re-merge in-place at four λ values bracketing the deployed PaSM mean ($\lambda \in \{0.32, 0.40, 0.55, 0.66\}$) and evaluate on GSM8K-100. PaSM's deployed selection on this pair concentrates near $\overline{\lambda}_{\text{attn}} = 0.34$, $\overline{\lambda}_{\text{ffn}} = 0.31$ (Table 19). Direct GSM8K-100 accuracy across the four sweep points is $\{0.81, 0.82, 0.67, 0.56\}$ at $\lambda \in \{0.32, 0.40, 0.55, 0.66\}$ respectively; these four points are from an additional in-place re-merge sweep, not from the cached grid of Table 15 (which samples the in-cliff region $\lambda \geq 0.50$). The base ($\lambda = 0$) GSM8K-100 accuracy is 0.85 from the cached grid. At the deployed operating point ($\lambda \approx 0.32$) the merged model retains 0.81 **accuracy, within 4 pp of base**; accuracy then falls -18 pp by $\lambda = 0.55$ and -29 pp by $\lambda = 0.66$, confirming the cliff location used by PaSM's validation argmax. Output verbosity drifts mildly at the operating point (output tokens 438–446 vs the base's ~ 150, hit-max-tokens rate 0.56–0.59), consistent with the Thinking-direction tuning bleeding into the merged checkpoint without collapsing the reasoning chain.

Scope of capability evidence. We do not claim full capability preservation across the entire trust action bank, nor that arbitrary weight-space interpolation points are capability-safe; Tables 15 and 16 together with the cliff measurements above themselves show that they are not. External capability probes are now available on *all four* paper pairs: GSM8K-100 scalar-λ sweeps on Qwen3-4B (operating-point bracketing), Phi-4 (7-point), Mutant (7-point bimodal), and Ministral (7-point rise-and-plateau). Capability behavior is pair-conditioned: Ministral is retention-favorable (PaSM-region GSM8K $\sim +4$ pp above base); Phi-4 GSM8K peaks at 0.83 at λ=0.80 with no capability-collapse region; Mutant is GSM8K-bimodal; Qwen3-4B is the boundary case with a generation cliff at $\lambda \approx 0.67$ that PaSM avoids. These are 100-query GSM8K probes, not global capability guarantees. The remaining 14B-class native LED/entropy-grid full paired-bootstrap/max-T aggregation on Phi-4 is reserved because its SNIP/backward pipeline has documented OOM risk. The present claim is bounded by these pair-specific sweeps, the selection-distribution statistics on all four pairs, and the deterministic plug-in-margin argument (Proposition), which applies to triples where the bank-oracle ceiling exceeds $2\eta(q)$; this condition is dense on the calibration axis where bank-oracle utility is high and the empirical $\eta(q)$ is small, while on the binary-utility truth and safety axes it reduces to the joint criterion $\max_a U_{ax}(q, a) = 1$ *and* $\eta(q) < 0.5$ that we report empirically rather than certify formally.

U Trust Card Worked Example: Extended Discussion

Table 3 in §5.5 shows a concrete Trust Card for the Ministral pair under balanced $\mathbf{w} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Truth and safety fall in the align-tune sparse-routability band (3–15% per-query headroom; §3.2): truth is served from the measured-bank fixed action, while safety admits the on-demand re-merge controller, which lifts the pair from $9/4$ to $18/8$ pb/max-T (§5.3). Calibration is dense-routable and uses the TextKNN-$L16$ specialist. The card therefore records the routing axis, deployed specialist, and materialization cost for serving.

V Native LED-Merging and Entropy-Grid In-Place Runs

The proxy baselines adapted to the same bank in Table 1 are diagnostic adaptations under the shared trust action bank evaluator. To check that the cell-count gap is not an artifact of running comparators inside our bank, we additionally implement LED-Merging (Ma et al., 2025) and an entropy-grid baseline (DaWin-inspired (Oh et al., 2025);

we batch $K = 5$ λ candidates and pick the per-query argmin of mean per-token entropy, which is one reasonable instantiation of the per-sample entropy-gated selection principle in DaWin but is not an exact reimplementation of Oh et al.'s entropy-ratio coefficient) as in-place per-prompt weight-composition pipelines. Native LED is run on all four paper pairs (full attention+FFN scope on Qwen3-4B, Ministral, and Mutant; attention-only reduced scope on Phi-4); entropy-grid is run on **Qwen3-4B, Ministral, and Mutant**. Phi-4 entropy-grid is reserved because the 14B-class native pipeline has documented OOM risk when combined with full statistical aggregation. LED-Merging uses per-prompt SNIP attribution ($|\nabla\theta{\cdot}\theta|$) on attention/FFN tensors and a disjoint top-K merge (safety-only ← intervention, utility-only ← base, mixed regions ← linear λ blend); the entropy-grid baseline batches $K = 5$ candidate λ points and picks the per-query argmin of mean per-token entropy.

Table 20 supports the proxy-control rankings in Table 1: native LED reaches comparable axis-utility to the bank-adapted proxies on each of the four pair types, with full attention+FFN scope on three pairs and an attention-only reduced scope on Phi-4 14B (footnote †). On Qwen3-4B and Ministral, LED's best calibration utility ranges 0.575–0.628 with safety reaching 0.83–0.87, and the truth values are comparable to the proxy-control rows in Table 1. On Mutant, LED reaches 0.67 truth, 0.75 safety, and 0.8089 calibration at the best per-axis configurations; the balanced s=0.20, u=0.10 configuration reaches 0.67/0.73/0.8089 on truth/safety/calibration. *The Mutant entropy-grid ($K = 5$) baseline collapses on the LoRA-merge pair class*, reaching 0.02 truth and 0.00 calibration; the entropy-only routing signal is structurally insufficient on this pair where most candidates carry similar entropy but vastly different per-axis utility. Similarly, entropy-grid loses substantial safety utility on Ministral (0.63) compared to LED (0.83), a per-axis routing failure consistent with our headroom analysis of §3.2. PaSM's headline cell-count metric on the 336-cell grid (Table 1) is reported under the paired-bootstrap and max-T family-wise correction; the native-implementation evidence here is a construct-validity check on the proxy adaptation, not a single-slice raw-utility comparison.

A side observation from the entropy-grid run is that the per-query λ selections themselves show clear per-axis structure on Qwen3-4B: 92/100 truthfulness queries pick the near-base $\lambda = 0.10$ point, 73/100 calibration queries pick the same low-λ region, while safety spreads across the full candidate set ($\lambda \in \{0.10, 0.30, 0.50, 0.70, 0.90\}$ with counts 27/24/20/14/15). Note that this pattern only partially aligns with the bank-oracle headroom asymmetry of §3.2: truth agrees (both diagnostics see little routing structure), but the entropy-grid's safety λ spread on Qwen3-4B is wider than the bank-oracle headroom would predict (§3.2 reports sparse truth+safety on the same pair). The two diagnostics measure different objects (predictive entropy across λ candidates versus bank-action utility headroom), and we report the λ-distribution as an independent observation rather than as corroboration of the main headroom claim. The remaining Phi-4 native-implementation

config	truth	safety	cal-util
Qwen3-4B (Instruct → Thinking)			
LED (s0.05/u0.20)	0.73	**0.87**	**0.628**
LED (s0.05/u0.30)	0.73	0.85	0.625
LED (s0.10/u0.20)	0.73	**0.87**	0.613
Entropy-Grid ($K = 5$ argmin entropy)	0.72	0.85	0.626
Ministral 8B (Base → Reasoning)			
LED (s0.20/u0.10)	**0.70**	**0.83**	0.575
LED (s0.10/u0.05)	0.68	0.82	0.577
LED (s0.20/u0.05)	0.69	0.79	**0.583**
Entropy-Grid ($K = 5$ argmin entropy)	0.71	0.63	0.587
Mutant 8B (Instruct → Mutant-Reasoning)			
	0.67	**0.75**	**0.8089**
LED top-3 per axis	0.67	0.74	0.8055
	0.65	0.73	0.8054
LED Pareto (s0.20/u0.10)	**0.67**	0.73	**0.8089**
Entropy-Grid ($K = 5$ argmin entropy)	0.02	0.64	0.00
Phi-4 14B (Base → Reasoning) – attention-only LED scope†			
LED (s0.20/u0.05)	0.83	0.50	0.66
LED Pareto (16 configs flat)	0.80	0.50	0.66

Table 20: Native LED-Merging on all four paper pairs and entropy-grid (DaWin-inspired) baselines on three pairs (Qwen3-4B Instruct→Thinking; Ministral 8B Base→Reasoning; Mutant 8B Instruct→Mutant-Reasoning; Phi-4 14B Base→Reasoning). Top LED results per pair shown: Qwen3-4B 15 configs swept; Ministral 16/16 configs swept; Mutant 16/16 configs swept; Phi-4 16/16 configs swept with attention-only reduced scope. The Mutant top-3 row lists per-axis utilities in descending order over the 16-config sweep. Calibration is the confidence-aware utility ($1-|\text{conf}-\text{correct}|$); truth and safety are parse-match accuracy on the same evaluator used throughout the paper. †Phi-4 LED was constrained to attention-only importance scope (skipping FFN gradients) due to 24GB-class GPU backward-OOM on the full attention+FFN scope; the sweep was also reduced to $n = 30$ per axis. The resulting utilities are nearly flat across the 16 (s, u) configurations (calibration $0.66 - 0.67$; safety locked at 0.50; truth $0.80 - 0.83$), i.e., reduced-scope LED on Phi-4 cannot differentiate between (s, u) configurations, consistent with the 14B-class XSTest safety ceiling and the broader pattern that external entropy-driven baselines saturate at larger scale. Full-scope LED on Phi-4 remains future work pending A100-class hardware or DeepSpeed ZeRO-3 offload.

Pool	**pb / max-T**	Δ
Nested-CV lower bound	$35/14$	$-19/-7$
Cached selection-only	$45/17$	$-9/-4$
Pre-declared (Mut+Min re-merge)	$\mathbf{54/21}$	**headline**

Table 21: Pool-sensitivity ladder. The headline is robust around the pre-declared pool: a later full-pool audit that also admits Qwen3-4B re-merge and Phi-4 re-merge nets $0/-1$ relative to the headline; bootstrap-seed robustness of the pre-declared pool across 8 seeds gives max-$T = 21$ across all 8 seeds and paired-bootstrap $\{52, 54, 54, 53, 54, 54, 55, 54\}$ (modal 54); a strict nested-CV protocol bounds the conservative anchor at $35/14$ (top row of the ladder).

comparison with paired-bootstrap CI and max-T correction is left for future work because of 14B-class SNIP/backward OOM risk on this pair.

W Re-merge Pool Sensitivity

The headline 54/21 (paired-bootstrap / max-T) is computed under the *pre-declared re-merge pool*, which admits the Mutant and Ministral on-demand re-merge controllers. The headline is robust under bootstrap-seed perturbation and a later full-pool admission audit; the strict nested-CV protocol is reported separately as a conservative lower bound ($35/14$), not as a robustness signal. Table 21 reports the compact pool-sensitivity ladder so reviewers can audit selector-pool sensitivity directly.

X Reproducibility Statement

All experiments use a fixed 5-fold cross-validation, paired-bootstrap CI ($B = 1000$ cluster bootstrap, cluster $= (\text{fold}, q)$), and Westfall–Young max-T ($B = 5000$ cluster sign-flip permutations, single-step) over the 84-cell axis $\times$ weight grid per pair. The primary selection-only 45 paired-bootstrap and 17 per-pair max-T cells (out of 336 audited), and the optional Ministral re-merge total of 54 and 21 cells, are computed from cached controller predictions and on-demand re-merge diagnostics; exploratory entropy-gated variants are reported as negative-ablation rows rather than headline claims. Upon acceptance, we will release a reproducibility package with source code, the trust action bank construction pipeline, hidden-state-cache metadata, controller-prediction files, and SHA256-verified scripts for regenerating the headline numbers, including a CPU-only regeneration path for the selection-only branch, the re-merge branch inclusion rule, and the proxy-baseline ladder.