WHEN ROUTING ACCURACY MISLEADS: AUDITING FRAMEWORK SELECTION WITH CERTIFIED TWINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many reasoning tasks require committing to a single epistemic framework before solving: blending a p-value with a posterior yields incoherence, not a partial truth. We study *framework routing* — selecting the applicable belief-space "universe" before reasoning — and report that high routing accuracy is a poor measure of it. Auditing our own benchmark, we find that a classifier that cannot read the words, only their shape and punctuation, reaches 76.1% against a 56.6% majority baseline, and that most ILL-POSED training items announced their own deficiency in a parenthetical present in 90.6% of train but 5.5% of test items. We release **EpistemicBench v5**, an explicitly versioned repair — v4 is untouched — that removes that marker together with the train-only augmentation tags, and discloses the leakage it does not fix. The repair works and does not buy the capability: retraining lifts ill-posed recall by $+47.9$pp, yet on a *certified* subset whose labels are machine-checkable properties of the generating parameters, our router detects $0/30$ ill-posed items, trained with the marker or without it. A pre-registered, unaudited replication over eight disjoint generator families repeats this at $0/100$. Items within a family are near-duplicates, so these are finite-suite counts, not binomial ceilings. Benchmark accuracy on this class can therefore be a severe overestimate of epistemic competence, and we make no claim that ill-posedness detection was learned. On the repaired data we detect no effect of gate hardness: every paired 95% CI among matched 3.7M hard, soft and Bayesian gates includes zero — a failure to detect a difference, not a demonstrated equivalence. A 1.6M fine-tuned specialist beats all three and ties RoBERTa-large at $1/220$ its trainable budget, but it differs from them in architecture and objective at once, so that comparison is observational. EpistemicBench v5, both certified sets, the audit tooling and code will be released upon publication. Audit the benchmark before believing the number.

1 INTRODUCTION

Consider an autonomous agent solving mathematical problems. **Q1**: "A coin flipped 100 times shows 60 heads. Is it fair ($\alpha = 0.05$)?" **Q2**: "Given uniform prior on bias θ, what is $P(\theta > 0.6 \mid 60$ heads)?" Both involve coin flips, yet Q1 demands **frequentist statistics** (null hypothesis testing and p-values), while Q2 demands **Bayesian statistics** (prior specification and posterior integration). These embody **incompatible epistemologies** (Jaynes, 2003; Berger, 1985): frequentists treat parameters as fixed unknowns; Bayesians treat them as random variables. An agent conflating these produces errors such as "the p-value is the probability the hypothesis is true," which is incorrect in kind rather than degree.

Scaling alone does not guarantee epistemic coherence (Kadavath et al., 2022; Yin et al., 2023): fluency and framework discipline are different properties. The problem is at least partly architectural — agents lack an explicit mechanism for recognizing *which framework* applies before solving. We formalize this as classifying questions into K **belief-space universes**: reasoning frameworks with distinct axioms and solvers (Def. 1, Figure 1). Our benchmark sets $K{=}12$ and instantiates them as 12 *operational routing categories* spanning frequentist and Bayesian statistics, classical/quantum/relativistic physics, Euclidean and non-Euclidean geometry, potential outcomes and structural causal models. We call them operational because they are not a clean partition of one kind: they mix framework (BAYESIAN, QUANTUM) with problem *status* (ILL-POSED) and with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

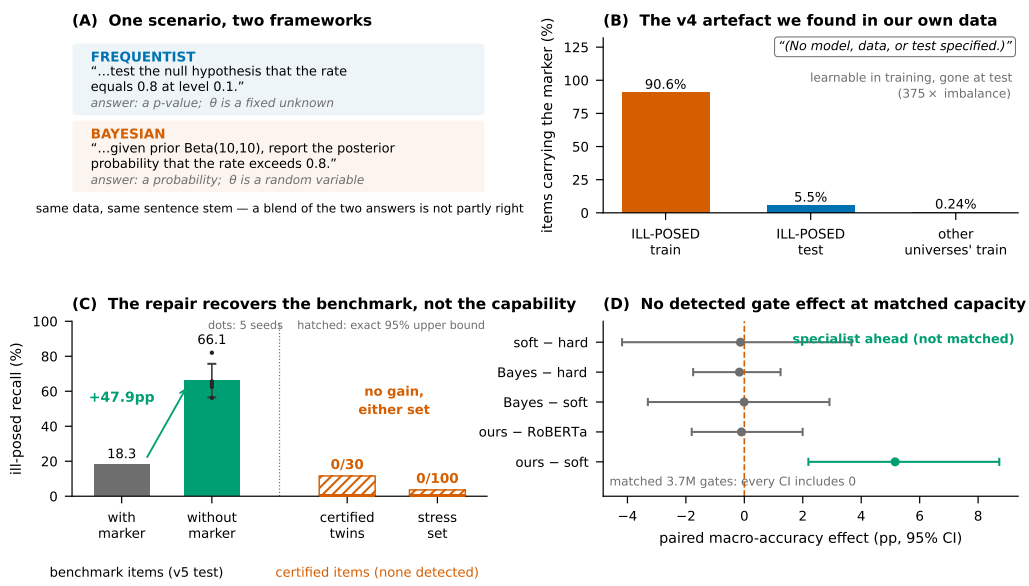


Figure 1: **What we set out to measure, and what we found we were measuring.** (A) Two questions over one dataset demand incompatible frameworks; a blend of a p-value and a posterior probability is not partly right. (B) In EpistemicBench v4, 90.6% of ILL-POSED *training* items announced their own deficiency in a parenthetical that appears in only 5.5% of its test items and 0.24% of other universes' training items — a $375\times$ imbalance a model can learn and then lose. (C) Removing it and the train-only tags (v5) raises ill-posed recall by $+47.9$pp on the benchmark, with wide seed spread (dots: five seeds, 56–82%). On *certified* items, whose labels are computed rather than annotated, detection is $0/30$ on the twins (two checkpoints) and $0/100$ on the pre-registered stress set (eight systems); hatching is the binomial 95% upper bound (11.6%, 3.6%), too narrow here because pairs within a family are near-duplicates (App. Y). Benchmark accuracy here can severely overestimate competence. (D) At matched 3.7M capacity over five retraining seeds we detect no difference among soft, hard and Bayesian gates (every paired 95% CI includes zero, which is not evidence of equivalence). The fine-tuned specialist is ahead of all three, but it also differs in architecture and budget, so that margin is observational.

cross-framework *relations* (CROSS-FRAMEWORK), so not all 12 are universes in the strict sense of Def. 1. We flag the mismatch here rather than only in Limitations because it shapes what the routing numbers can mean.

This paper is mostly about why that framing is harder to evaluate than it looks. We built the benchmark and trained the router, then audited our own data and found that a classifier which cannot read — one that maps every letter to `a` and every digit to `0`, preserving only punctuation and shape — very nearly matches it. One universe's training items announced their own labels in a parenthetical; four universes are produced by the same generator on both sides of a supposedly held-out split. None of this is visible in an accuracy number.

We repaired what could be repaired and versioned it explicitly, and we built an evaluation whose labels are computed rather than annotated: minimal *twin* pairs differing by one clause, one member provably non-identifiable and the other provably determined, each carrying a machine-checkable certificate recomputed by independent code. The certificate binds the generating parameters, not the prose — that each rendered question states exactly its certified constraints is author-checked, not independently audited. On that set our routers detect $0/30$ ill-posed items — the repair that bought $+47.9$pp on the benchmark bought nothing here. We think this gap, not any accuracy number, is the paper's contribution.

Contributions. (1) ***A benchmark audit and an explicitly versioned repair.*** We search our own labels for surface features that predict the universe without carrying epistemic content, and report what we find: a self-announcing deficiency marker in 90.6% of ILL-POSED train items versus 5.5%

of test items; a content-blind classifier at 76.1%; the source corpus predicting 78.7% of labels; and four universes sharing generator provenance across the split, which is why we rename the evaluation *source-ID-disjoint* (§4.1). EpistemicBench v5 removes the marker, ships a manifest, checksums and the unfixed findings, and leaves v4 unmodified so prior numbers remain reproducible. **(2)** ***Certified twins show benchmark accuracy can severely overestimate competence.*** Repairing the marker lifts ill-posed recall by $+47.9/+55.4$pp across two architectures — so the "worst-group blind spot" we previously reported was a property of the text, not of the models. But on the certified twins our router, trained with and without the marker (one checkpoint each), detects $0/30$ ill-posed items, gets both twins right in $0/30$ pairs, and gives both twins one label in 25–28 of 30 pairs; removing framework vocabulary while preserving framework semantics drops a router from 100% to 23% pair accuracy (§4.2). We claim no learned ill-posedness detection. **(3)** ***No detected effect of gate hardness.*** At matched 3.7M capacity over five retraining seeds, every paired contrast among hard, soft and Bayesian gates has a 95% CI including zero (soft $-$ hard $= -0.16$pp $[-1.21, +0.82]$), on v5 and on all three frozen-metadata subsets — a failure to detect a difference on this benchmark, not a demonstrated equivalence. A 1.6M-parameter fine-tuned specialist outperforms the matched gates and ties fully fine-tuned RoBERTa-large at $1/220$ the trainable budget, but differs from them in architecture and objective too, so that comparison is observational (§4.3). Two propositions (§3) motivate why incompatible *outputs* resist averaging; we present them as diagnostic criteria, not as claims any routing number verifies.

Downstream, conditioning a fixed solver on the routed framework changes its terminology but *reduces* answer accuracy, a negative result we quantify with injection-blind scoring (§4.5). Calibrated uncertainty, continual learning, robustness and MMLU validation are in App. B.

2 Related Work

LLM-based Agents. ReAct (Yao et al., 2023) interleaves reasoning and actions; Reflexion (Shinn et al., 2023) adds self-correction via verbal reinforcement; chain-of-thought prompting (Wei et al., 2022) improves multi-step reasoning. Recent surveys (Huang & Chang, 2023; Tao et al., 2024; Wang et al., 2024b) identify lifelong learning without catastrophic forgetting (Kirkpatrick et al., 2017; De Lange et al., 2022; Wu et al., 2024) as a central challenge. These works improve execution *within* a reasoning procedure; we address the upstream problem of *selecting* the framework.

Classifier-based LLM Routing. A growing line of work trains small classifiers to route queries among LLMs. RouteLLM (Ong et al., 2025) learns BERT-scale routers from preference data to dispatch between strong and weak models; Hybrid LLM (Ding et al., 2024) routes for cost–quality tradeoffs; FrugalGPT (Chen et al., 2023) cascades models with learned scoring; Shnitzer et al. (2023) route across model zoos using benchmark statistics; concurrent work also routes by adaptive cost (Vahidi, 2026), by interaction state for multi-agent collaboration (Wang et al., 2025a), or hybridizes soft and hard routing over LoRA-expert pools (Huang et al., 2026). Mixture-of-Agents (Wang et al., 2024a) aggregates outputs of multiple LLMs through a peer-aggregator architecture, a representative "soft combination of LLMs" baseline. These methods all route or combine on *cost or expected quality* of otherwise interchangeable models, which share an underlying solution space. We route on *epistemic compatibility*: experts encode mutually contradictory axioms, so the choice is not which model is cheaper or stronger but which framework is even applicable — the condition under which mixing incurs an interior penalty (Prop. 2). The distinction is also empirical: retraining a RouteLLM-scale encoder classifier (BERT-base, LoRA) directly on our 12-universe labels reaches only $90.64 \pm 0.87\%$ over 3 seeds on v4 — 3.9–5.6pp below our router, significantly in every seed (exact McNemar, Holm-corrected for multiple tests) — so that architecture class does not close the gap when repurposed for epistemic routing (App. C).

Mixture of Experts: routing, failures, and Bayesian variants. Classical MoE (Shazeer et al., 2017; Fedus et al., 2022) and Mixtral (Jiang et al., 2024) softly combine homogeneous experts; recent compute-aware extensions train MoE under domain uncertainty via bilevel sparse optimization (Chen et al., 2025). Failure-mode analyses include representational limits of soft combination (Chung et al., 2024), token-level gradient conflicts (Yang et al., 2024b; Gan & Lei, 2025), sparse-MoE collapse (Do et al., 2025), and graph-based stabilization (Nguyen-Nhat et al., 2025). Bayesian routing variants (Li et al., 2025; Li & Wicker, 2026) improve soft routing in the *mixture-representable* setting, where the target lies inside the convex hull of the components and an interior

mixture beats every single expert; Prop. 2 concerns the complementary *vertex-identifiable* regime (Bernardo & Smith, 2009), where the target coincides with one component and is not reachable by blending the others. Concurrent domain-decoupled work (Innovator's fine-grained MoE upcycling (Liao et al., 2025)) shares the empirical intuition. Closest to our empirical setup, Molavi et al. (2026) find that *hard* top-1 routing over frozen reasoning-LoRA experts matches or beats soft mixtures at a fraction of the trainable budget, and report soft top-2 gates concentrating most mass on one expert (dominant weight ≈ 0.71). Their experts are trained on *mutually compatible* benchmarks (GSM8K, ARC, MedQA, BoolQ, CoLA) and selected on cost/quality grounds. We reimplement their straight-through-estimator (STE) mechanism at matched 3.7M capacity (§4.3) and detect no hardness effect: on v5, every paired 95% CI among the STE hard gate, a soft top-2 gate and an MC-dropout Bayesian gate includes zero over five seeds. Our experiments therefore separate two questions that routing work often conflates — whether *specialization* helps, and whether the *gate* must be hard. On the second we detect no effect on this benchmark. On the first our evidence is observational: a fine-tuned specialist outperforms the gates, but it also differs from them in architecture and objective.

Gradient geometry, continual learning, and uncertainty. Subspace orthogonality bounds catastrophic forgetting (Steele, 2026); SETA (Siddika et al., 2026) and SEE (Wang et al., 2025b) build mixture-of-experts and continual-learning methods on this insight, the latter sidestepping the router by dispatching on out-of-distribution scores. We extend deep ensembles (Lakshminarayanan et al., 2017; Kendall & Gal, 2017) and split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021; Romano et al., 2020) to epistemic routing (App. B).

3 METHOD

3.1 PROBLEM FORMULATION

Definition 1 (Belief Space Universe)**.** *A belief-space universe $u \in \mathcal{U}$ is a reasoning framework (A_u, I_u, S_u): axioms A_u, inference procedure I_u, and solver S_u. Two universes u_i, u_j are epistemically incompatible if $\exists\, a \in A_{u_i}, b \in A_{u_j}$* with a common syntactic referent *(i.e., a and b predicate over the same variable, event, or quantity) such that $b \Leftrightarrow \neg a$.*

Operational scope. The shared-referent restriction in Def. 1 places our notion closer to Carnap's framework relativization than to Kuhnian incommensurability (Kuhn, 1962; Feyerabend, 1975; Davidson, 1974): universes share the observable space (the question q) and incompatibility is local to predicates with overlapping syntactic content (e.g., both quantify the same probability $P(\theta)$). Universes are therefore *operationally* incommensurable, not globally so. The restriction rules out vacuous incompatibility between any two non-overlapping axiom systems.

Definition 2 (Universe Routing)**.** *Let $\mathcal{Q}$ be the set of questions and $P_{\mathcal{Q}}$ a distribution over it. Given $q \sim P_{\mathcal{Q}}$, the routing function $r_\theta : \mathcal{Q} \to \mathcal{U}$ assigns q to universe $u^* = \arg\max_{u \in \mathcal{U}} P(u \mid q; \theta)$, where θ denotes the router's learned parameters throughout. The full system computes $S_{u^*}(q)$.*

Proposition 1 (Non-certifiability of Contradictory Provenance)**.** *Assume each solver emits an output together with a derivation recording the propositions it invokes (*proof provenance*), and that an output is* certified *under universe u when a consistent axiom set A_u contains its provenance. Suppose a combined output $\hat{y}$ has provenance containing both a proposition a and its negation $\neg a$. Then no consistent single-universe axiom system certifies $\hat{y}$.*

Epistemically incompatible expert pairs (Def. 1) provide a mechanism by which such contradictory provenance can arise, when the conflicting axioms are active in the two derivations being combined. Incompatibility alone does not imply that every mixed output invokes them: two derivations may draw only on the fragment their axiom sets share. The proof is in Appendix K, with three worked examples in Appendix M. We claim non-certifiability, not that the *value* of $\hat{y}$ can never coincide with a separately-valid answer.

In words, the condition below asks that *the right answer be a corner, not a blend*: no weighted average of the *other* experts reproduces it. Where one does, averaging wins and the proposition does not apply — that is the regime soft and Bayesian mixtures are built for, and our main scope limit.

Proposition 2 (Interior-Mixture Penalty). *Let $p_{\mathrm{mix}}(y|x) = \sum_k w_k\, p_k(y|x)$ and suppose the correct component $k^*(x)$ is identifiable with $P^*(\cdot|x) = p_{k^*}(\cdot|x)$. Assume the* vertex-identifiability *condition: $p_{k^*}(\cdot|x)$ is not representable by the remaining* indexed *components,*

$$p_{k^*}(\cdot|x) \;\notin\; \mathrm{conv}\{\, p_k(\cdot|x) : k \neq k^* \,\}. \tag{1}$$

Then under any strictly proper loss the population risk is uniquely minimized by the vertex weight $w = e_{k^*(x)}$: any gate assigning $w_{k^*}(x) < 1$ on a set of positive P_X-measure incurs positive *population excess risk (deviation on a null set changes nothing). Input-dependent soft gating $w(x)$ attains it only in the limit where its mass concentrates on $k^*(x)$ — vanishing softmax temperature is one route there, and only when $k^*(x)$ already carries the largest logit. We use Prop. 2 as a diagnostic regime boundary, not a claim that no soft scheme can ever match hard routing.*

Why this condition. Strict propriety makes $p_{k^*}(\cdot|x)$ the unique risk-minimizing *predictive distribution* but does not pin the *weight vector*. Pairwise disagreement is insufficient: a non-degenerate mixture of components that each differ from p_{k^*} can equal it (e.g., $p_{k^*} = \frac{1}{2}p_2 + \frac{1}{2}p_3$). Condition (1) rules this out and is necessary as well as sufficient. It is stated over the indexed set because a duplicated component would otherwise break uniqueness (App. L). Without it the proposition weakens to "the risk-optimal predictive distribution is p_{k^*}". Epistemic incompatibility motivates this regime but is not a premise: the result applies to any vertex-identifiable family, and incompatibility does not imply it.

Scope of the propositions. Both propositions concern the *output space*: the non-certifiability of combined outputs with contradictory provenance (Prop. 1) and the interior-mixture penalty for combining predictive distributions (Prop. 2). Our routing experiments (§4) instead measure the accuracy of a 12-way classification *gate* — a related but distinct object, and one our own results show can be driven by surface structure rather than by the semantics the propositions describe (§4.1–4.2). We therefore present the propositions as criteria that *motivate* hard commitment at the output level, not as theorems any routing-accuracy number verifies.

3.2 Dataset Construction

We construct **EpistemicBench**, a 22,009-sample benchmark spanning 12 universes across 5 domains, drawn from 12 upstream corpora recorded under 125 source identifiers, most of them subject- or split-level aliases (48 MMLU subjects (Hendrycks et al., 2021), Hendrycks MATH, UGPhysics, SciQ (Welbl et al., 2017), ARC (Clark et al., 2018), TheoremQA, MedQA, among others); 85.6% of items come from these corpora and 14.4% are synthetic augmentation for rare universes. Rule-based classification assigns labels by domain-specific keyword matching with per-universe validation; every universe has $\geq$500 samples. We name the twelve labels in small capitals; App. O gives the released label strings. The original v4 split (19,661 train / 906 val / 1,442 test, with four stratified diagnostic subsets defined in Appendix O) is used only for legacy v4 comparisons; all primary results use the v5 source-ID-disjoint split of §4. EpistemicBench will be released upon publication under the license matrix of the Ethics Statement.

3.3 Model Architecture and Training

We fine-tune Qwen-1.5-0.5B (the Qwen-1.5 release of the Qwen family (Bai et al., 2023; Yang et al., 2024a); 464M non-embedding / 620M total parameters; we round the non-embedding count to 465M throughout) with low-rank adaptation (LoRA; rank 16, $\alpha = 32$) on the query and value projections, and a 12-class classification head. A deep ensemble of 5 models differing only in random seed provides uncertainty estimates (Lakshminarayanan et al., 2017). Training on 19,661 samples takes $\sim$90 minutes on a single RTX 3090.

Only LoRA adapters and the 12-class head are trainable (1.6M parameters: 1.58M LoRA on the query and value projections of all 24 layers, plus a 12K-parameter head); the backbone is frozen. Pooling, head architecture, optimizer settings, full hyperparameters and seeds are in Appendix T.

4 EXPERIMENTS

Our evaluation is ordered by what it licenses: we audit the benchmark and repair what we find (§4.1), ask whether the repaired version measures epistemic competence (§4.2) — it does not — and only then compare routers on it (§4.3–§4.5).

Metrics and split. Micro-accuracy is over items, macro-accuracy averages the twelve per-class recalls, and worst-group is the lowest of them. All primary results use a *source-ID-disjoint* split (14,006 train / 6,447 test, no source corpus appears on both sides; held out: Hendrycks MATH geometry/precalculus/counting, SciQ and MMLU-Pro physics, our synthetic generator, and MMLU philosophy/moral disputes). We say source-*ID*-disjoint, not source-disjoint, because the guarantee is on the source *identifier* and not on the data-generating process: four universes — BAYESIAN, MIXED/PHILOSOPHICAL, ILL-POSED and STRUCTURAL-CAUSAL — are 100% generator-produced on *both* sides, so for them "held out" names a different string, not a different process.

4.1 AUDITING THE BENCHMARK: WHAT PREDICTS A UNIVERSE LABEL?

Before comparing routers we asked what a model *could* exploit. We trained classifiers on six families of surface features that carry no epistemic content, the last of them the content-blind classifier of §1, which keeps only punctuation, casing and length (App. V). The result is uncomfortable: on the repaired split that classifier alone reaches 76.1% micro-accuracy against a 56.6% majority baseline, character affixes 70.3%, punctuation 62.6%. The metadata is no better — naming the source corpus alone predicts 78.7% of training labels under a majority rule, and STRUCTURAL-CAUSAL's test items sit 89% behind three prefixes.

The audit also found one outright artifact. Most ILL-POSED training items announced their own deficiency in a parenthetical — *"(No model, data, or test specified.)"* — present in $308/340$ of its train items (90.6%) against $33/13{,}666$ of every other universe's (0.24%, a $375\times$ imbalance), but in only 5.5% of test items, so a model can key on it during training and lose it at test. Removing it, and train-only augmentation tags of the form *"(Variant 2)"*, uniformly across all twelve universes and retraining over five seeds raises ILL-POSED recall by $+47.9$pp $[+32.6, +60.8]$ (ours) and $+55.4$pp $[+46.2, +64.4]$ (RoBERTa-large), and worst-group accuracy by $+36.8$ and $+40.0$pp (bootstrap means, which for worst-group do not equal the difference of the seed means; Table 20). The marker and the tags come out together, so these are the effect of the combined v4→v5 edit, not of the marker alone. Both architectures move together, which is what an artifact predicts and a model-specific weakness does not.

The repair ships as **EpistemicBench v5**, explicitly versioned with a manifest and per-file checksums; **v4 is unmodified**, so previously published numbers remain reproducible and are retained only as the artifact ablation. We stress what v5 does *not* fix: the surface-classifier figures above are measured *on v5*, after the repair. Removing a marker we could name does not remove the structural leakage we could only measure, and v5 ships that finding in its list of known limitations. Nor did the repair catch every variant of the marker. Four rarer deficiency parentheticals (e.g., *"(The sample space is not defined.)"*) that its pattern did not match remain on 32 of 340 ILL-POSED training items (9.4%) and on no other item in any split. We found them after v5 was frozen, and because releases are never edited in place, we disclose them here rather than patch them.

4.2 CERTIFIED TWINS: BENCHMARK GENERALIZATION IS NOT EPISTEMIC RECOGNITION

A benchmark whose labels come from keyword rules cannot distinguish a router that learned the epistemology from one that recovered the rules. We therefore built a *certified* subset in which the label follows from the generating parameters and carries a machine-checkable certificate, so no annotation — human or model — assigns it. The certificate covers the mathematics, not the rendering: that each question states exactly the constraints it assumes is author-checked, and the labels are relative to the frequentist/Bayesian taxonomy of App. A, under which a likelihoodist reading would move some. The certified set removes the annotator from the label, not every judgment from the task. Items are emitted only as **minimal pairs**: two questions sharing a scenario and differing by one clause, across the label boundary.

Composition. The suite is 50 pairs (100 items): 20 FREQUENTIST/BAYESIAN pairs, and the 30 ILL-POSED/well-posed twins this section's numbers refer to, from five templates that each withhold a different quantity (base rate, sample size, prior, stopping rule, α). Neither family was filtered after seeing results. Every $0/30$ is over those 30, the whole ill-posed family and not a selection from it.

An item is ill-posed *iff* two admissible completions of the unstated quantity satisfy every stated constraint yet disagree by a margin fixed per template before generation (0.20 in probability, 0.02 in p-value, or — where the question asks for a decision — a flip in the reject/retain conclusion itself). *The certificate is that pair of completions*, recomputed from the generator's parameters by a second implementation that imports none of the generator's code (390 checks, zero failures). The twin restates the identical scenario with the quantity supplied, so its answer is unique. Overlap within a pair is measured, not assumed: mean token Jaccard 0.61, min 0.35.

> **Ill-posed.** "A screening test has sensitivity 0.98 and specificity 0.9. A person drawn from the screened population tests positive. What is the probability that they actually have the condition?"
> *Certificate:* prevalence is free; $0.001 \Rightarrow 0.0097$, $0.1 \Rightarrow 0.521$; divergence $0.51 >$ this template's 0.2.
>
> **Well-posed twin.** "…and the condition affects 0.002 of the screened population. …"
> *Certificate:* the withheld quantity is supplied, so the completion set is a singleton.

Neither checkpoint detects anything. We evaluate our router trained with the marker and trained without it, one checkpoint each (seed 1; these are not seed averages). Both detect $\mathbf{0/30}$ certified ill-posed items (binomial 95% upper bound 11.6%, assuming independence). Neither gets *both* members of any ill-posed pair right ($0/30$ pair accuracy; TF-IDF likewise $0/30$). Instead they assign the *same* label to both twins in $25/30$ ill-posed pairs (83%, exact 95% CI $[65, 94]$; marker-free) and $28/30$ (93%, $[78, 99]$; marker-trained). Marker removal, which bought $+47.9$pp on the benchmark class, buys nothing here and makes the well-posed twins *worse* ($12/30$ vs. $24/30$). The 30 pairs come from five templates, so they describe these generators and these two checkpoints.

A cue ablation isolates the mechanism. On the 100-pair Family A set, the production router (trained on v4, one checkpoint) is perfect when framework vocabulary is present ("null hypothesis", "significance level", "prior Beta$(10, 10)$"): 100% item, 100% pair. Rewriting the same questions to preserve the epistemic content while removing that vocabulary — "if the defect rate were exactly 0.8, how often would 40 units produce a result at least as extreme" versus "a density proportional to $x^9(1-x)^9$" — collapses it to 57.0% item $[50.1, 63.7]$ and 23% pair $[15.8, 32.2]$ (Wilson binomial 95% intervals) (exact McNemar b=86, c=0, $p<10^{-25}$). The failure is strongly one-sided: recall is $83/100$ $[74.5, 89.1]$ on the frequentist member and $31/100$ $[22.8, 40.6]$ on the Bayesian one (Wilson), while precision on BAYESIAN stays $31/31 = 100\%$ (exact 95% lower bound 88.8%), so the router is not defaulting to a majority class but losing a class entirely. Pair accuracy (23%) is close to the 25.7% expected under independent failures, so we do not claim correlated within-pair errors (Fisher one-sided p=0.10). Benchmark accuracy on these classes can therefore be a *severe overestimate* of epistemic competence — an empirical finding about this benchmark and these routers; we prove no bound.

4.3 SPECIALIST ROUTING AND MATCHED GATES

No detected gate effect. Within the matched 3.7M mixture-of-experts family we retrain a soft top-2 gate, a Monte-Carlo-dropout Bayesian gate and the straight-through hard top-1 gate of Molavi et al. (2026) over five seeds, pair them symmetrically (seed-i vs. seed-i), and estimate each effect by a hierarchical bootstrap over seeds *and* samples. Every paired effect is a fraction of a point with a 95% CI spanning zero: soft − hard $= -0.16$pp $[-1.21, +0.82]$, Bayesian − hard $= +0.12$pp $[-0.59, +0.81]$, Bayesian − soft $= +0.29$pp $[-0.18, +0.78]$ on micro-accuracy, and likewise on macro. We therefore detect no effect of gate hardness, and we do not claim the gates are equivalent: no equivalence margin was pre-specified, and while every micro-accuracy interval lies within ± 1.21pp, the macro intervals are wider (soft − hard -0.13pp $[-4.18, +3.67]$), so macro effects of a few points are not excluded (App. X).

The specialist is ahead; the comparison is observational. Our 1.6M fine-tuned specialist outperforms the 3.7M matched gates by $+1.9$pp micro $[+1.0, +2.7]$ and $+5.2$pp macro $[+2.2, +8.7]$ against soft, and $+1.7$/$+5.0$pp against hard, both with CIs excluding zero, on fewer trainable pa-

Table 1: **Primary results.** EpistemicBench v5 (repaired), source-ID-disjoint split, n=6,447; five retraining seeds per neural row (mean ± sample SD, $\mathrm{ddof}=1$ throughout), TF-IDF one deterministic run. The **shaded block** is the only capacity-matched comparison: one 3.7M MoE backbone, gate swapped, hard = straight-through (Molavi et al., 2026); no metric separates the three. The last column is the seed-paired bootstrap against the hard gate, not the difference of the Macro column. Only the specialist's $+5.0$pp excludes zero, and it changes architecture, objective and budget at once, so that margin is observational.

Method	Trainable	Micro	Macro	Worst-group	Δ Macro vs. hard
Reference systems: not capacity-matched, not seed-paired					
TF-IDF + logistic regression	—	82.7%	72.1	39.8	—
RoBERTa-large (full fine-tune)	355M	87.9±0.8%	80.0±1.8	57.8±6.2	—
Matched gates: same 3.7M backbone, only the gate changes					
soft top-2	3.7M	86.7±1.0%	74.7±4.0	50.5±8.6	-0.1 $[-4.2, +3.7]$
Bayesian, MC-dropout	3.7M	87.0±0.6%	74.7±1.3	51.4±6.7	-0.2 $[-1.8, +1.2]$
hard top-1, straight-through	3.7M	86.8±1.0%	74.9±1.1	51.2±4.0	*reference*
Specialist: no gate, one fine-tuned classifier					
Qwen-1.5-0.5B + LoRA (ours)	**1.6M**	88.6±0.1%	79.9±1.2	56.9±7.0	$\mathbf{+5.0}$ $[\mathbf{+3.8}, \mathbf{+6.4}]$

rameters. Against fully fine-tuned RoBERTa-large (355M, 220× the trainable budget) it is a tie: -0.10pp macro $[-1.80, +2.00]$. Our router is a classification head, *not* a capacity-matched gate: it differs from the gates in architecture, training objective and trainable budget at once, so this margin is not a controlled estimate of what specialization contributes.

Sensitivity. Because the audit showed the split is confounded, we recompute every matched-gate effect on three subsets defined *only* by metadata frozen inside v5 — degenerate templates, synthetic sources, universe membership — no model output enters their definition (App. X). No matched-gate contrast is detected on any of them. The specialist's margin does not survive equally well. Dropping the four shared-generator universes roughly halves its macro margin over the matched gates ($+5.2/+5.0 \rightarrow +2.3/+2.3$pp, CIs $[+0.5, +4.7]$ and $[+1.0, +3.9]$), so part of that advantage is attributable to universes whose train and test halves come from one generator. On real items alone, its margin over the soft gate is no longer detected ($+2.0$pp macro $[-0.5, +4.8]$; $+0.7$pp micro $[-0.1, +1.2]$), while the margin over the hard gate remains ($+2.8$pp macro $[+0.2, +6.4]$).

4.4 A Disjoint Certified Set: Semantic Counterfactual Stress

Pre-registered; a diagnostic, not a confirmatory result. To test whether §4.2's $0/30$ is a template artifact, we pre-registered **SCS-1**, a semantic counterfactual stress set: 300 minimal pairs from eight disjoint generator families, certificates independently recomputed and frozen before any prediction. No item carries the marker of §4.1: ill-posedness must be inferred from content. Eight systems — five seeds, the marker-trained checkpoint, two lexical baselines — detect $\mathbf{0/100}$ certified ill-posed items and give both twins one status in $\mathbf{300/300}$ pairs, while the same five checkpoints recall 56–82% of ill-posed items on the v5 test split. The binomial bound over those 100 items (3.6%) is too narrow to read as a ceiling: predictions are deterministic within family × phrasing, so there are at most sixteen independent observations. Under the pre-registered family-level cluster bootstrap none is distinguishable from zero. Its protocol required a blind expert audit; that did not happen, so no claim here rests on it (App. Y).

The items are decidable; the router is what fails. Asked only whether a question as stated carries enough information to have one answer — one prompt fixed before the run, every output parsed — Claude Sonnet 5 scores $100/100$ on SCS-1, recall, specificity and pair accuracy alike, and $24/30$ pair accuracy on §4.2's twins; a local Qwen2.5-7B-Instruct recalls $100/100$ at $75/100$ specificity. Both are handed the question our gate must infer and answer a binary form of it, so this is not a like-for-like comparison. What it settles is the objection that matters: $0/100$ and $0/30$ are properties of the router, not of items that nothing could decide.

4.5 End-to-End Solver Conditioning

Routing is only useful if it changes what the solver does. We fix the solver to Qwen2.5-7B-Instruct, vary only the system prompt over 240 stratified questions — bare question, framework name, name plus axioms — and score how far its answer uses that framework's vocabulary while *ignoring any term the prompt itself supplied*, since the axioms condition injects 3.0 such terms per question and a raw score would reward echoing. We call this the *blind* conformance score.

Name-only conditioning raises blind conformance from 0.153 to 0.208 ($+36\%$) while *lowering* answer accuracy from 0.429 to 0.345; name+axioms inflates *raw* conformance to 0.453 but collapses the *blind* score to 0.033 — blind scoring withholds credit for every term the prompt supplied, so it cannot separate echoing from correct adoption. The reading is narrow: conditioning changes the solver's terminology, and we have no evidence it changes the reasoning. Conformance is a lexical proxy, and the one accuracy number we have moves the wrong way. Full table and protocol: App. S.1.

5 Discussion

What the results support. Three findings are durable. First, *routing accuracy on a rule-labeled benchmark is not a measurement of epistemic competence* — a failure of construct validity (Cronbach & Meehl, 1955). We report this because we found it in our own data: a router near ceiling on in-distribution ill-posed items scores $0/30$ on certified ones. We do not know how far it generalizes, but the mechanism is not special to us: keyword-derived labels can create shortcuts that let a router recover the annotation rule rather than the intended concept, and an accuracy number will not reveal it. Second, *we detect no effect of gate hardness* at matched capacity — a failure to detect a difference among three implementations on one benchmark, not evidence that gates are interchangeable in general, and the macro intervals still admit effects of a few points. Separately, a fine-tuned 1.6M specialist outperforms all three and ties a 355M fully fine-tuned encoder; because it also differs from them in architecture and objective, this suggests, but does not establish, that fine-tuning a specialist is a better use of budget here than a bespoke hard-routing mechanism. Third, the propositions of §3 say when output-level mixing is ill-posed; they are criteria that motivate commitment, not results any routing number verifies, and Prop. 2 holds only under the vertex-identifiability condition we state.

What the results do not support. The matched gates are not shown equivalent, only undistinguished. The specialist's margin is not evidence about gate hardness: that comparison changes architecture and budget too. There is no downstream gain — conditioning a fixed solver shifts its vocabulary but *reduces* answer accuracy (§4.5). Our routers detected nothing on either certified set, which is no evidence that recognition was learned rather than proof that it was not — and a statement about these checkpoints, not about routers in general. And the source-ID-disjoint split does not isolate semantic from lexical routing — our own audit refutes that, as it refutes the gradient-subspace geometry (App. J) as a diagnostic of incompatibility.

Limitations

Four scope limits bound our claims (App. A). (i) *Hard boundaries*: one universe per question, so within-question transitions fall to coarse buckets. (ii) *Normatively-contested domains* (ethics, contested law) are out of scope. (iii) *Labels and taxonomy*: labels are rule-based, the 12 universes are *operational labels*, not universes under Def. 1, and v5 repairs two artifacts but not the structural leakage or four universes' shared generator provenance (§4.1). A blind human audit of the MIXED/PHILOSOPHICAL/ILL-POSED boundary is in progress and unreported here; its pre-registered order was not kept — an AI assistant opened the answer key first. (iv) *Mixture-representable regimes*: where the target lies inside the components' convex hull, Prop. 2's premise fails and soft or Bayesian MoE should dominate, and we have no validated test for which regime an expert pool is in — the largest open problem this work leaves.

ETHICS STATEMENT

Licensing. EpistemicBench aggregates items from 12 upstream corpora (MMLU, MMLU-Pro, Hendrycks MATH, SciQ, ARC, TheoremQA, MedQA, UGPhysics, among others; 125 source identifiers) alongside rule-templated and generator-produced items we authored. We license *our own* contributions — the universe labels, provenance and artifact metadata, the v5 manifest, the template-generated and certified items, and all audit and evaluation tooling — under CC-BY-4.0. **Source items remain under their original terms**, which we do not relicense: aggregation confers no such right. What is redistributed is controlled by a per-identifier license matrix (125 identifiers, 14 families, each with verbatim license evidence). Only GSM8K and our template-generated items (together 11.9% of items) have clear redistribution rights. For the other ten upstream corpora (85.5%), no data license settles redistribution, the license is non-commercial, or the questions derive from third-party material the dataset authors do not own. For these we release identifiers and a reconstruction script, not text, and reconstruction depends on upstream availability: the canonical Hugging Face copy of Hendrycks MATH has been disabled following a DMCA notice. The 566 items from our synthetic generator, partly generated with OpenAI models, are withheld until the applicable terms are confirmed.

Scope. The routing technique targets reasoning-framework selection and is *not* intended for normatively contested domains such as ethics or contested law (see Limitations), where "which framework applies" is a political question rather than a technical one. **Evaluation.** All evaluations use public models and report parameters, latency and cost. We also note that this paper's central finding — that our own benchmark is substantially predictable from surface form — is a caution we apply to ourselves before others: released accuracy numbers on EpistemicBench should be read alongside the released list of known limitations and the certified twins, not on their own.

REPRODUCIBILITY STATEMENT

Model, pooling, optimizer settings, hyperparameters and all random seeds are in App. T; the dataset construction protocol is in App. O, and the artifact audit and the v4→v5 repair in App. V. EpistemicBench v5 is specified by a manifest with per-file SHA-256 checksums and a list of known limitations recording the leakage the repair does not remove; v4 is retained unmodified. Figure 1's panels B–D and the primary tables (Tables 1 and 22, and the per-universe breakdown) are generated by scripts directly from stored per-seed predictions, and the bootstrap procedure (seeds and samples resampled jointly, B=2000, a 30-item class-support floor) is described with them. Every certified item carries its certificate, which independent code recomputes (§4.2). Complete proofs of both propositions, with their assumptions stated explicitly, are in Apps. K and L. The algebraic core of Props. 1–2 is mechanically checked in Lean 4, and independent SMT/FOL checks are provided with the artifact. The protocol of the in-progress label audit, including its deviations, is in App. A.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. *Synthetic dataset generation:* OpenAI GPT-4o-mini and GPT-4 generated part of the synthetic items in EpistemicBench (a subset of one synthetic source, which appears only in the test split); the rest of that source, and all of the augmented synthetic items, are filled from templates by deterministic code. The certified twins and the stress-set items are produced by deterministic generator code, not by a language model. *Mathematical claim formulation and proof assistance:* an AI coding assistant (Claude Code) helped restate both propositions, write their proofs, and formalize their algebraic core in Lean 4 with SMT and first-order checks. The Lean kernel, not the assistant, is the verifier of the formal results. *Research methodology and experiment design, implementation, data cleaning, and results interpretation:* the same assistant helped design and implement the benchmark audit, the v5 repair, the certified evaluation, the sensitivity analyses and the statistical procedures, and helped interpret their results. *Conceptual framing and hypothesis refinement:* the same assistant helped develop the framing of both propositions — including the vertex-identifiability condition of Prop. 2, its converse, and the scope paragraph separating the propositions' output-space claims from what a routing-accuracy number can verify — and helped refine the hypotheses the audit and the certified sets were built to test. *Review and revision:* AI agents ran simulated expert-review and non-

specialist-reading passes over the manuscript, and the revisions those prompted include substantive corrections rather than wording alone. One identified an incorrect necessity claim in Prop. 2 — that the optimal vertex is attained *only* as the softmax temperature vanishes, which is false at fixed temperature with a diverging logit margin — and App. L now records it as an error of an earlier draft. Others found a statistical claim that contradicted our own appendix, an arithmetic inconsistency between two tables reporting the same condition, and a miscount of the surface-feature families in our own audit; each is corrected in the text. *Qualitative analysis:* both label sets collected so far for the taxonomy audit (App. A) were produced by AI assistants, one of them a second coding agent (Codex) used to audit the repository; neither counts as a human annotation, and no audit result is reported. *Separately from assistance*, language models are objects of study in this paper — routers, baselines, solvers, a generalist comparator on the certified sets (App. Y), and an automated label cross-audit (App. H) — and are reported as such. We also used generative AI tools for tasks with recommended disclosure: code creation and editing, figure creation (plotting scripts), and paper drafting and editing. We did *not* use AI tools for literature search or analysis, for research-topic identification, or to translate author-written text.

The authors read the proofs of both propositions and the generated benchmark items. Beyond that reading our verification is mechanical rather than manual: reported numbers are regenerated from stored result files by scripts rather than transcribed, certificates are recomputed by independent code that does not import the generator, and the formal proofs are checked by the Lean kernel. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.

José M Bernardo and Adrian FM Smith. Bayesian theory. *John Wiley & Sons*, 2009.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Xian Rui Chen, Bei Ling Gao, Yu Qiao Fang, An Zhen Liu, and Zhou Shen. Distributionally-robust gradient routing: A bilevel sparse optimization problem for compute-aware mixture-of-experts training. *Preprint, doi:10.21203/rs.3.rs-8381882/v1*, 2025.

Youngseog Chung, Dhruv Malik, Jeff Schneider, Yuanzhi Li, and Aarti Singh. Beyond parameter count: Implicit bias in soft mixture of experts. In *NeurIPS ENLSP Workshop*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Donald Davidson. On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47:5–20, 1974.

Matthias De Lange, Rahaf Aljundi, Marc Masana, et al. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V.S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations (ICLR)*, 2024.

Giang Do, Hung Le, and Truyen Tran. SimSMoE: Toward efficient training mixture of experts via solving representational collapse. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. doi: 10.18653/v1/2025.findings-naacl.107.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Paul Feyerabend. *Against Method*. Verso, 1975.

Yuxing Gan and Ziyu Lei. Mixture-of-experts with gradient conflict-driven subspace topology pruning for emergent modularity. *arXiv preprint arXiv:2512.20291*, 2025.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *Findings of ACL*, 2023.

Yitong Huang, Ziqi Yang, Zihui Wang, Jianzhong Qi, Rongshan Yu, et al. Hybrid routing for a mixture of LoRA experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. doi:10.1609/aaai.v40i37.40383.

Edwin T Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Albus Yizhuo Li and Matthew Wicker. Variational routing: A scalable Bayesian framework for calibrated MoE transformers. *arXiv preprint arXiv:2603.09453*, 2026.

Albus Yizhuo Li et al. Bayesian mixture-of-experts: Towards making LLMs know what they don't know. *arXiv preprint arXiv:2509.23830*, 2025.

Ning Liao, Xiaoxing Wang, Zehao Lin, Weiyang Guo, Feng Hong, et al. Innovator: Scientific continued pretraining with fine-grained MoE upcycling. *arXiv preprint arXiv:2507.18671*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Seyed Alireza Molavi, Zhan Su, Yan Hu, Peyman Sheikholharam Mashhadi, Stefan Byttner, and Prayag Tiwari. Learning to select, not relearn: Hard-routed mixtures of reasoning LoRAs. *arXiv preprint arXiv:2606.31413*, 2026.

Minh-Khoi Nguyen-Nhat, Rachel S. Y. Teo, Laziz U. Abdullaev, Maurice Mok, Hoang Tran-Viet, et al. Modeling expert interactions in sparse mixture of experts via graph structures. *arXiv preprint arXiv:2510.16411*, 2025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Noah Shinn, Federico Cassano, Ashwin Gopinath, et al. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. In *NeurIPS Workshop on Instruction Tuning*, 2023.

Fatema Siddika, Md Anwar Hossen, Tanwi Mallick, and Ali Jannesari. Split-on-share: Mixture of sparse experts for task-agnostic continual learning. *arXiv preprint arXiv:2601.17616*, 2026.

Brady Steele. Subspace geometry governs catastrophic forgetting in low-rank adaptation. *arXiv preprint arXiv:2603.02224*, 2026.

Zhengwei Tao et al. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*, 2024.

Soroush Vahidi. When to revise: Cost-aware adaptive routing for LLM reasoning. *Preprint, doi:10.2139/ssrn.6510975*, 2026.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, et al. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *Advances in Neural Information Processing Systems*, 35, 2022.

Jingbo Wang, Sendong Zhao, Hao Wang, Yuzhen Fan, Lizhe Zhang, et al. Optimal-agent-selection: State-aware routing framework for efficient multi-agent collaboration. *arXiv preprint arXiv:2511.02200*, 2025a.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a.

Lei Wang, Chen Ma, Xueyang Feng, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024b.

Zhilin Wang, Yafu Li, Xiaoye Qu, and Yu Cheng. SEE: Continual fine-tuning with sequential ensemble of experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025b. doi: 10.18653/v1/2025.findings-acl.387.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Longrong Yang, Dong Shen, Chaoxiang Cai, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Solving token gradient conflict in mixture-of-experts for large vision-language model. *arXiv preprint arXiv:2406.19905*, 2024b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *Findings of ACL*, 2023.

Appendix Contents

A Limitations: Full Discussion **16**

B Supporting Analyses: Uncertainty, Robustness, Continual Learning, External Validation **17**

C Batch 1 v4-Unified Architecture and Fair-MoE Numbers **17**

D Conditional Control: A Non-Epistemic Expert Pool **17**

E Gating Entropy vs. Accuracy Across Matched Variants **17**

F Few-Shot 2026 Micro LLMs: Full Sweep **18**

G Fair-MoE Gating Sweep on the Hard Split **18**

H Automated Cross-Audit of v4 Labels **20**

I EpistemicBench v5: Per-Universe Breakdown **21**

J Exploratory Gradient Geometry **22**

K Proof of Proposition 1: Non-Certifiability of Contradictory Provenance **23**

L Proof of Proposition 2: The Interior-Mixture Penalty **23**

M Error Propagation: Empirical Demonstration **25**

N Cloud / Generalist-LLM Evaluation Details **25**

N.1 Few-shot 7–8B LLMs and Mixture-of-Agents (v4 stratified-360) 25

N.2 v4 stratified-360 zero-shot (apples-to-apples) 25

N.3 Pilot n=109 cloud generalists (legacy summary) 26

O Dataset Construction Protocol **27**

P Continual Learning: Extended Analysis **27**

P.1 Replay Buffer Sensitivity (8 ratios $\times$ 3 seeds) 27

P.2 Expansion Order Robustness 28

Q External Validation: MMLU Details **28**

R Worked Case Study: Double-Slit and the Behavioral Effect of Routing **28**

S Adversarial Robustness, Uncertainty, Conformal Prediction, Scaling, Ablations: Tables **29**

S.1 End-to-End Solver Conditioning: Full Table 29

S.2 Adversarial Robustness 29

S.3 Deep Ensemble Uncertainty Quantification 29

S.4 Conformal Prediction: Standard vs. Variance-Scaled Scores 30

S.5 Ablations and Calibration Details 30

T Reproducibility **30**

U Full Routing Comparison (v4-Unified) **31**

V Benchmark Artifact Audit and the v4→v5 Repair **31**

W Legacy v4 Specialist-vs-Routing-Variant Comparison **33**

X Sensitivity of the Gating Null to Benchmark Artefacts **34**

Y SCS-1: Semantic Counterfactual Stress Set **34**

Each appendix is self-contained and is referenced from the main text at the point where it is used. No step of the main text's argument depends on reading one.

A LIMITATIONS: FULL DISCUSSION

This expands the four scope limits summarized in the Limitations section.

(i) Hard-boundary scope. Universe routing assumes one universe per question; problems requiring within-question framework transitions (a quantum-scattering question whose classical limit is taken; a Bayesian–frequentist asymptotic-agreement argument) fall to the STAT_MIXED_OR_PHILOSOPHY or CROSS_FRAMEWORK buckets, which is a coarse mitigation. Principled top-k routing from a conformal prediction set is left to future work.

(i-b) Two statistical frameworks, not three. Our statistical partition is frequentist versus Bayesian. It omits *likelihoodism* — the position that evidence is carried by the likelihood ratio alone, with neither long-run error rates nor a prior — which is a third foundational stance, not a blend of the two. Its absence has a concrete consequence for §4.2: several of the certified ill-posed items withhold a prior or a stopping rule, and a likelihoodist would regard some of those as answerable from the likelihood alone rather than as underdetermined. Our certificates are therefore certificates *relative to* the two frameworks the taxonomy encodes, and the $0/30$ and $0/100$ results should be read that way. Adding a likelihoodist universe would change the label of some items and is left to future work.

(ii) Normatively-contested domains. Routing where frameworks are themselves political or ethical positions (normative ethics, contested law) risks reifying ideological priors and is out of scope.

(iii) Benchmark labels and taxonomy. EpistemicBench v4 is 85% real data and 15% template-augmented for rare universes. Rule-based labels carry $\sim$5% noise on well-defined universes by a 250-sample human spot-check, but an automated LLM cross-audit (App. H, n=600) agrees with the rules only 63.7% overall, 26–42% on philosophically subtle universes, and *0%* on STAT_ILL_POSED. That audit *cannot* adjudicate correctness: without independent human ground truth, LLM–rule disagreement is equally consistent with LLM error and with label error. It is therefore neither a validation nor a bound in either direction — only an indicator of where ambiguity concentrates. An audit by an independent human annotator is the outstanding step for benchmark credibility. It is in progress, and we describe its protocol together with its deviations as they stand. It targets the STAT_MIXED_OR_PHILOSOPHY/STAT_ILL_POSED boundary over 48 shuffled items in three length-matched strata — 16 *case* items that a majority of retraining seeds reassign, 16 STAT_MIXED controls and 16 STAT_ILL_POSED controls — with three labels (MIXED / ILL_POSED / neither). The items, instructions, answer key, and three of four decision branches were fixed before any labeling. *Deviations.* (a) The first label set was produced by the AI assistant that built the v5 repair, and it had already seen gold labels or model votes for 7 of the 48 items. The plan therefore treats the other 41 as primary. (b) The written analysis plan (raw agreement, Cohen's κ over all 48 and over the 41, per-stratum agreement, and the overlap of "neither" labels against its independence expectation with a one-sided Fisher test) was committed shortly after that label set. Its fourth branch, which blocks any local relabelling if both annotators place a nontrivial common subset outside the binary, was added after seeing the first label distribution. It has no numeric threshold, so any branch outcome will be reported as a disclosed judgment. (c) The registered order — the key opened only after two independent label sets are committed — was broken. A second AI assistant, auditing the repository, filled the second annotation sheet and then opened the answer key, before any independent human had labeled. Its labels are not an independent annotation and are excluded from the confirmatory analysis. The analysis plan predates that key access and is unchanged, but the independent human annotation now being collected is recorded on a separate sheet and does not restore the original order. *No result from this audit appears in this paper*, and no conclusion here depends on it. Relatedly, the 12 universes are *operational routing labels* mixing framework (STAT_BAYES, PHYS_QUANTUM), problem status (STAT_ILL_POSED), and multi-framework relations (CROSS_FRAMEWORK); not all are belief-space universes in the strict sense of Def. 1, and a hierarchical domain/framework/status scheme would be cleaner. The standard split shares sources between train and test (TF-IDF 95.8%); the source-ID-disjoint split (§4.3) controls for this.

(iv) Mixture-representable regimes. When the target conditional lies in the *interior* of the convex hull of the candidate components — multilingual translation experts, intent-classification specialists, arbitrary code-domain experts — the vertex-identifiability condition of Prop. 2 fails, and Bayesian MoE (Li et al., 2025; Li & Wicker, 2026) or standard soft MoE should dominate. Our reading of Prop. 2 is therefore conditional on the regime, not unconditional. We stress that we have no validated *test* for which regime a given expert pool is in: the gradient-angle statistic of App. J is exploratory

and confounded with topical separation, and our own certified evaluation (§4.2) shows that a router can score highly on a benchmark built for one regime while failing the controlled diagnostic entirely. Diagnosing the regime from data is open.

B Supporting Analyses: Uncertainty, Robustness, Continual Learning, External Validation

Four supporting analyses are summarized here; full protocols and tables follow in the appendices cited. **(i) Uncertainty.** A 5-seed deep ensemble matches the best single seed in accuracy, and its mutual information is $10.7\times$ higher on errors, separating them with AUROC 0.94 (App. S.3) — discrimination, not calibration. Scaling the conformal nonconformity score by that mutual information improves coverage at a mean prediction-set cost under 0.07 classes (App. S.4). **(ii) Robustness.** The fine-tuned router is far less brittle than the lexical baseline to keyword-injection, synonym, and mixed-language attacks (overall attack success 1.5% vs. TF-IDF's 65.8%; App. S.2). **(iii) Continual learning.** Rehearsal with $\sim$10% replay brings forgetting on a $5\to7$ universe expansion down to $9.9 \pm 7.2\%$ — below the 10% line, not near zero, and with a seed spread as wide as the effect (App. P). **(iv) External weak-label validation.** On 1,001 MMLU questions the router beats TF-IDF by $+10.6$pp under identical noisy proxy labels (App. Q). Routing errors propagate benignly: with p_r=0.954 and solver accuracy 0.95, end-to-end accuracy is 0.911 ± 0.006, within 0.1pp of the analytic bound $p_e \leq p_r p_s + (1 - p_r)/K$.

C Batch 1 v4-Unified Architecture and Fair-MoE Numbers

The v4-unified runs supplementing Table 18 and Sec. 3 are summarized in Table 2. All trained with FP32, 5 epochs, batch size 16, AdamW lr 2×10^{-4} (LoRA) / 1×10^{-3} (head), label-smoothing 0.05; LoRA r=16 on W_Q, W_V. Note that BERT-base ($90.64 \pm 0.87\%$) underperforms DistilBERT (92.16%) under matched-LoRA training—a known LoRA-rank-vs-backbone interaction; we use the Qwen-1.5-0.5B production router for headline numbers.

D Conditional Control: A Non-Epistemic Expert Pool

On a non-epistemic control task (7-category 20 Newsgroups, matched size and architecture), soft MoE gating entropy is 0.047, against 0.000 on the epistemic baseline. We report the difference and stop there.

Earlier drafts read it as evidence that strong gating commitment is *selectively* a property of epistemically incompatible expert pools — a validated signature of the regime. We withdraw that reading, for three reasons. Both values are near zero and the comparison is between two datasets that differ in far more than epistemic compatibility (class count, domain, text length, label noise). Entropy does not predict accuracy across the 20 matched runs (Spearman ρ=0.15, p=0.53; App. E), so even a real entropy difference would not license an accuracy claim. And the matched-gate comparison on v5 is a null (§4.3): if commitment were the operative variable, the hard gate should have separated from the soft one, and it does not. We therefore report a numerical difference on this control and do not interpret it as a diagnostic for which regime an expert pool occupies. Finding such a diagnostic remains open.

E Gating Entropy vs. Accuracy Across Matched Variants

All five matched routing variants collapse to near-degenerate routing ($H \leq 0.15$), and within that family entropy does not predict accuracy (Spearman ρ=0.15, p=0.53), so collapse alone does not determine accuracy (Figure 2).

Table 2: Batch 1 v4-unified architecture comparison and Fair-MoE row, all on EpistemicBench v4 (n=1,442). Matched soft/stochastic and STE rows report 3-seed means±std (the seed-paired analysis of Table 21); BERT-base is the 3-seed RouteLLM-scale classifier of §4; remaining rows are single-run. n_e = number of experts.

Model	Trainable	Total	Acc.
Architecture (LoRA r=16, head)			
DistilBERT-base	895K	67M	92.16%
BERT-base	599K	110M	90.64±0.87%
Qwen-1.5-0.5B (LoRA narrow)	1.59M	465M	94.66%
Qwen-1.5-0.5B (production)	1.6M	465M	95.42%
Fair MoE (3.7M soft)			
Fair MoE (n_e=4, top-2)	3.7M	465M	93.50±0.56%
seed-mean gating entropy = 0.064 (§4.3)			
Bayesian variants (re-run on v4)			
Bayesian Temperature MoE (unmatched)	—	—	83.98%
Bayes. Temp.-Stoch. (matched, K=8)	3.7M	465M	93.57±0.42%
Bayes. Logit-Gauss. (matched, K=8)	3.7M	465M	93.62±0.55%
Bayes. MC-Dropout (matched, K=8)	3.7M	465M	93.64±0.58%
STE hard top-1 (Molavi et al., 2026)			
STE hard (v4 standard split)	3.7M	465M	93.78±0.49%
STE hard (source-ID-disjoint hard)	3.7M	465M	88.48%
Cross-base (2026-vintage)			
Qwen3-0.6B (production LoRA r=16)	4.4M	600M	94.73%

F Few-Shot 2026 Micro LLMs: Full Sweep

Table 3 reports the full 0/3-shot sweep that Table 8 compresses to one row per model. All under greedy (T=0) decoding with 4-bit NF4 quantization; Qwen3 family uses `enable_thinking=False`.

Table 3: Full zero- and 3-shot sweep of the four 2026 micro LLMs on the 360-sample stratified v4 subset.

Model	Params	Shots/u	Accuracy
Qwen3-0.6B	0.6B	0	12.50%
Qwen3-0.6B	0.6B	3	19.17%
Qwen3.5-2B	2B	0	40.56%
Qwen3.5-2B	2B	3	55.56%
Gemma-3-1B-it	1B	0	19.72%
Gemma-3-1B-it	1B	3	31.67%
Gemma-4-E2B-it	2B (E)	0	69.44%
Gemma-4-E2B-it	2B (E)	3	64.44%

G Fair-MoE Gating Sweep on the Hard Split

What this sweep does and does not test. It checks that the soft gate's hyperparameters were not badly chosen: we sweep the gating softmax temperature $\tau \in \{0.5, 1.0, 2.0\}$ and the top-k count

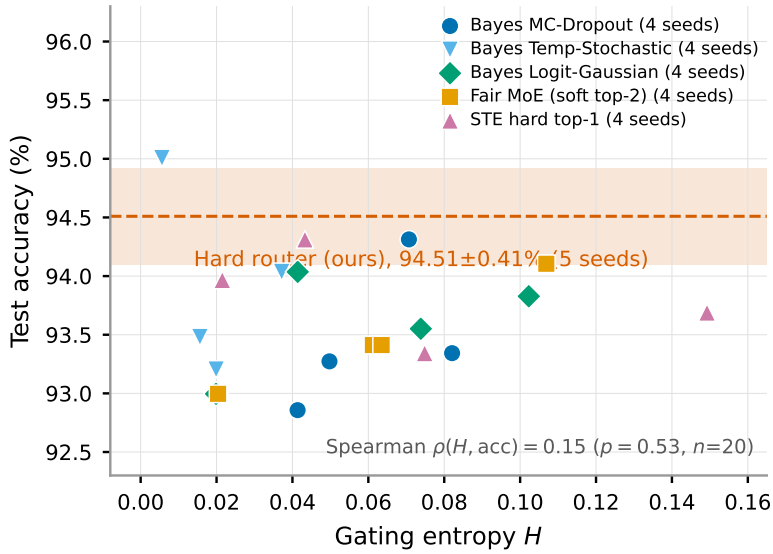


Figure 2: Accuracy vs. gating entropy for all five matched routing variants — three Bayesian, one soft top-2, and the STE *hard* top-1 gate, so this is not a soft-only family (3.7M trainable each, 2.3× the hard router's 1.6M; v4, n=1,442), 4 training seeds per variant (20 runs). The dashed line and shaded band are the 1.6M specialist's own 5-seed retraining distribution ($94.51 \pm 0.41\%$; Table 21), *not* the single 95.42% deterministic checkpoint, so the seed dispersion is like-for-like. The comparison to that band is seed-symmetric but *not* architecture- or capacity-matched: the variants carry 3.7M trainable parameters against the specialist's 1.6M. Every gate collapses toward near-degenerate routing ($H \leq 0.15$); within this family, H does not predict accuracy (Spearman ρ=0.15, p=0.53) — collapse alone does not determine accuracy. Bayesian gates marginalize over K=8 passes; STE follows Molavi et al. (2026).

$k \in \{1, 2, 4\}$ while holding the rest of the recipe identical (LoRA r=16 on $q_{\text{proj}}, v_{\text{proj}}$, four expert heads of width $\frac{1}{2}\cdot$hidden, AdamW lr 5×10^{-5}, 5 epochs, batch size 16, 3.7M trainable parameters per configuration). It is *not* a hard-vs-soft test: the comparison row is our 1.6M classification-head specialist, which differs from the 3.7M MoE variants in architecture and trainable budget as well as gate, and every configuration here is a single run. The controlled three-way gate comparison is the five-seed matched-capacity experiment on v5 (§4.3), and it is null. Read this table as evidence about the *specialist-vs-MoE* gap, consistent with the +1.7–1.9pp margin reported there — a margin that confounds architecture, objective and budget, and that we do not call a specialization effect.

Table 4: Fair-MoE gating-hyperparameter sweep on the source-ID-disjoint hard split (single run per configuration). All five configurations trail our 1.6M specialist (88.0%) by 0.66–2.88pp. The best soft configuration (τ=2.0, k=2) reaches 87.34%; sharper gating (τ=0.5) hurts. This shows the soft gate was not crippled by its hyperparameters; it does not isolate a hardness effect.

τ	top-k	n_e	Acc.	Gate H
0.5	2	4	85.12%	0.064
1.0	1	4	86.96%	0.670
1.0	2	4	86.97%	0.388
1.0	4	4	86.92%	0.133
2.0	**2**	4	**87.34%**	0.408
Hard router (ours)			**87.99%**	—

Discussion. Three patterns emerge. (i) Sharper gating (τ=0.5, entropy collapse to 0.06) *reduces* accuracy by 1.85pp, so the residual gap to the specialist is not explained by how close the soft gate sits to one-hot — if anything, pushing the soft gate toward hardness hurts. (ii) The best configuration uses *softer* gating (τ=2.0) yet still trails by 0.66pp, a difference well inside the between-seed disper-

sion measured in §4.3 and not separately significant here. (iii) The top-k variants at τ=1.0 are tied (86.92–86.97%) across entropies of 0.13–0.67, so accuracy is insensitive to top-k once temperature is fixed. Together these say the gating hyperparameters are not the operative variable. Combined with the null in §4.3, that makes gate hardness an unlikely explanation of the specialist's margin — but it does not identify what does explain it. The specialist differs from the MoE variants in architecture, training objective and trainable budget simultaneously, so we do not attribute the margin to fine-tuning, or to any single one of the three; ruling out one candidate is not isolating another.

H AUTOMATED CROSS-AUDIT OF V4 LABELS

We conduct a secondary automated audit of EpistemicBench v4 labels: 50 stratified samples per universe (600 total) are independently labeled by Qwen2.5-7B-Instruct (4-bit NF4, T=0 greedy decoding) using the same 12-universe definitions and the prompt format of App. N. We compute the machine-vs-rule agreement rate per universe and the dominant confusion pairs. *This is not a substitute for human annotation*: an LLM annotator shares many biases with our rule-based pipeline (both rely on surface tokens), so a low agreement rate indicates ambiguity or LLM-annotator failure as much as rule-based labeling error.

Table 5: Automated machine-vs-rule audit on 600 stratified v4 samples (50 per universe). Overall agreement 63.7%. The dominant disagreement is `STAT_ILL_POSED` (0% agreement; the LLM annotator routes ill-posed questions to `STAT_FREQ`), which is the same phenomenon the paper identifies in App. N for LLMs as solvers. By classification method, agreement is human-annotated 55.8%, rule-based 51.3%, template-generated 77.0%.

Universe	n	Agree	Dominant LLM error
STAT_BAYES	50	100%	—
CAUSAL_SCM	50	98%	1 to MATH
MATH_NON_EUCLIDEAN	50	90%	5 to MATH_EUCLIDEAN
PHYS_CLASSICAL	50	86%	QUANTUM/RELATIVITY
STAT_FREQ	50	84%	STAT_MIXED
PHYS_RELATIVITY	50	80%	8 to PHYS_CLASSICAL
STAT_MIXED_OR_PHILOSOPHY	50	68%	11 to STAT_BAYES
PHYS_QUANTUM	50	52%	23 to PHYS_CLASSICAL
MATH_EUCLIDEAN	50	42%	16 to NON_EUCLIDEAN, 7 to STAT
CAUSAL_PO	50	38%	13 to STAT_MIXED, 9 to STAT_FREQ
CROSS_FRAMEWORK	50	26%	31 to STAT_MIXED
STAT_ILL_POSED	50	**0%**	**36 to STAT_FREQ, 8 to STAT_MIXED**
Overall	600	**63.7%**	—

Interpretation. The audit surfaces three distinct phenomena rather than a single "noise rate":

1. **Ambiguous case: LLM blindspot or label defect.** `STAT_ILL_POSED` 0% agreement coincides with the failure mode the paper reports for LLMs as *solvers* (App. N, Table 9), which *suggests* an annotator blindspot. But we cannot conclude this: using an LLM to audit labels the same LLM family cannot recognize is circular, and 0% agreement is equally consistent with the rule labels themselves being wrong. Without independent human ground truth the two explanations are indistinguishable; we do not claim the audit validates the `STAT_ILL_POSED` labels, and treat the ill-posedness results as provisional (see Limitations).
2. **Inherently philosophical universes.** `CROSS_FRAMEWORK` (26%) and `STAT_MIXED_OR_PHILOSOPHY` (68%) are defined by spanning multiple frameworks;

LLM annotators reasonably disagree on which framework dominates. We acknowledge this in Limitations.

3. **Potential rule-based labeling noise.** MATH_EUCLIDEAN (42%) and CAUSAL_POTENTIAL_OUTCOMES (38%) have lower agreement consistent with both LLM error and possible rule-pipeline imprecision. Template-generated samples agree more (77.0%) than rule-based ones (51.3%). That gap is consistent with rule-based labels for real-data universes being noisier than the ∼5% spot-check estimate, but it is equally consistent with template items simply being easier for an LLM annotator to classify, so we do not convert it into a noise figure in either direction.

We previously described this automated audit as a sanity check that labels are not catastrophically noisy on the well-defined universes. We withdraw that reading. It is not a validation of anything: an LLM–rule disagreement is equally consistent with LLM error and with label error, and the audit reports 0% agreement on STAT_ILL_POSED — the class the certified evaluation (§4.2) later showed no router can detect at all. It is neither a bound on label noise nor a validation of the labels — only an indicator of where ambiguity concentrates. An independent human audit is the outstanding step for the taxonomy's credibility; one is in progress on the STAT_MIXED/STAT_ILL_POSED boundary and is not reported here (protocol and its deviations: App. A, item iii).

I EpistemicBench v5: Per-Universe Breakdown

Table 6 reports per-universe recall on **v5**, the same 6,447-sample source-ID-disjoint test set used for the primary results (§4.3), so the aggregate row reconciles with Table 1 rather than with the superseded v4 numbers. Columns are TF-IDF + LogReg (one deterministic run), fully fine-tuned RoBERTa-large (355M trainable) and our Qwen-1.5-0.5B + LoRA router (1.6M trainable), the latter two as mean±SD over five retraining seeds. The table is produced by aggregating the per-sample prediction dumps of the v5 chain; no model is retrained for it.

Table 6: Per-universe recall on EpistemicBench v5, source-ID-disjoint split. Δ_{R-T} is our router's gain over TF-IDF, sorted descending. Two readings are worth separating. The two *repaired* classes move most — STAT_ILL_POSED +26.2pp and STAT_MIXED_OR_PHILOSOPHY +28.0pp over the lexical baseline — but §4.2 shows the ill-posed gain does not transfer to certified items, so it should not be read as epistemic recognition. The three negative rows are stress-test failures where the held-out source style was absent from training, chiefly CAUSAL_POTENTIAL_OUTCOMES (test draws from MMLU-prof_med_test/BBH-causal_judgement, train mostly template_fill_v3). CAUSAL_SCM is at ceiling for every method, which App. V explains: 89% of its test items sit behind three prefixes. The two encoders tie per-universe; our router's aggregate micro edge over RoBERTa-large is a MATH_EUCLIDEAN composition effect (57% of the split).

Universe	n	TF-IDF	RoBERTa-L	Ours	Δ_{R-T}
STAT_MIXED_OR_PHILOSOPHY	132	45.5%	75.9±2.1	73.5±5.7	+28.0pp
STAT_ILL_POSED	128	39.8%	71.9±7.1	66.1±9.6	+26.2pp
STAT_BAYES	238	84.9%	92.8±5.3	97.3±0.8	+12.4pp
STAT_FREQ	405	54.3%	61.4±10.3	64.7±3.6	+10.4pp
PHYS_QUANTUM	168	77.4%	88.7±1.2	86.7±1.5	+9.3pp
CROSS_FRAMEWORK	805	89.6%	97.5±0.3	98.0±0.3	+8.4pp
PHYS_RELATIVITY	78	64.1%	66.7±4.2	69.0±5.7	+4.9pp
MATH_EUCLIDEAN	3,651	88.6%	92.3±0.6	93.0±0.6	+4.4pp
PHYS_CLASSICAL	373	86.1%	86.1±0.9	89.4±0.2	+3.3pp
CAUSAL_SCM	74	100.0%	100.0±0.0	98.4±3.6	-1.6pp
MATH_NON_EUCLIDEAN	151	63.6%	62.4±0.6	58.7±7.8	-4.9pp
CAUSAL_POTENTIAL_OUTCOMES	244	70.9%	64.5±4.1	64.2±2.9	-6.7pp

Discussion. Source-ID-disjoint evaluation is harsher than the standard split and shows that rare-universe transfer — CAUSAL_POTENTIAL_OUTCOMES, MATH_NON_EUCLIDEAN — remains un-

resolved. Both the lexical baseline and the strong supervised baseline drop on the same universes, so the limitation is shared across approaches rather than specific to our recipe. We treat these numbers as a stress test, not a deployment estimate, and — given the audit — not as a measure of semantic routing either.

J Exploratory Gradient Geometry

Status. This analysis is exploratory and supports no claim in the paper. Earlier drafts stated it as a proposition and used a 45° angle as a diagnostic threshold for epistemic incompatibility; we have withdrawn both. Three reasons. (i) The forgetting law $\mathcal{F} = c_1 \cos^2 \theta_{\min} + c_2$ is imported from Steele (2026) and empirically fit, not derived here. (ii) The 45° cut is the midpoint of that law's range, not a calibrated decision boundary; with 61 cross-domain and 5 intra-domain pairs it is fit and evaluated on the same data. (iii) Most importantly, our pairs are *cross-domain* proxies for incompatibility — statistics versus physics versus geometry — which need not share a syntactic referent in the sense of Def. 1, so a large $\theta_{\min}$ is equally consistent with topical separation. Since the certified evaluation of §4.2 provides a controlled semantic diagnostic that this angle statistic does not, we report the geometry below as an observation about the trained adapter and draw no design recommendation from it.

For each universe u, we compute loss gradients of the *production router* (router_v3) on 30 held-out samples per universe in its LoRA-delta space ($\Delta W = BA$ on each layer's W_Q, W_V projections; 1.6M parameters total across 24 layers). We stack the gradients into a matrix and extract the top-20 principal components by SVD. Pairwise minimum principal angles $\theta_{\min}$ between any two universes' subspaces are computed via the canonical-angles algorithm of Golub & Van Loan (2013) (seed 42). The underlying pretrained weights are frozen and contribute zero rank to the trainable subspace. All angles below are computed in the production LoRA-delta space (the fine-tuned adapter update); we note that in the *base* pre-fine-tuning weights even compatible same-domain pairs exceed 45°, so a trained adapter is required for the diagnostic to separate pairs at all — itself a limitation of the diagnostic (it cannot be evaluated before training).

Table 7: Minimum principal angles $\theta_{\min}$ (degrees) between gradient subspaces in the production LoRA-delta space, by pair category. Categories are *cross-domain* and *intra-domain*, which are topical groupings; we do not claim they are epistemically incompatible and compatible respectively. The 45° column is reported for comparability with Steele (2026), not as a threshold we endorse.

Pair Category	**Mean $\theta_{\min}$**	**vs. 45°**
Cross-domain (61 pairs)	64.5°	59/61 above
Intra-domain (5 pairs)	35.7°	4/5 below
PHYS_CLASSICAL↔PHYS_QUANTUM	19.9°	below
CAUSAL_PO↔CAUSAL_SCM	35.1°	below
PHYS_RELATIVITY↔PHYS_QUANTUM	36.9°	below
PHYS_CLASSICAL↔PHYS_RELATIVITY	39.5°	below
MATH_EUCLIDEAN↔MATH_NON_EUCLIDEAN	46.9°	above

The one intra-domain pair above 45° — MATH_EUCLIDEAN↔MATH_NON_EUCLIDEAN at 46.9° — is the pair whose axioms genuinely contradict (the parallel postulate and its negation), unlike the classical/relativistic/quantum pairs, which share a limiting regime. We note this because it is the one place where the geometry tracks Def. 1 rather than topic, but a single pair is an anecdote, not evidence: with 5 intra-domain pairs total, no statistic computed here separates the two explanations. Gradient subspaces of a *trained* adapter are more orthogonal across domains than within them, which is unsurprising, and that we cannot show this reflects epistemic incompatibility rather than topical distance.

K Proof of Proposition 1: Non-Certifiability of Contradictory Provenance

Let $\hat{y}$ have provenance Π with $a \in \Pi$ and $\neg a \in \Pi$, and suppose some consistent axiom set A_u certifies $\hat{y}$, i.e., $\Pi \subseteq A_u$. Then $a \in A_u$ and $\neg a \in A_u$, so A_u derives both a and $\neg a$ and is inconsistent, a contradiction. Hence no consistent single-universe axiom system certifies $\hat{y}$. □

What the proof does not use. Nothing in the argument refers to the experts that produced $\hat{y}$. In particular it does not use epistemic incompatibility (Def. 1). Incompatibility is the *mechanism* we have in mind for how contradictory provenance arises — a derivation under u_i invoking $a \in A_{u_i}$ combined with one under u_j invoking $b \Leftrightarrow \neg a$, $b \in A_{u_j}$ — but it does not guarantee that a given question activates the conflicting pair: both derivations may draw only on axioms the two universes share, in which case the combination has consistent provenance and the proposition says nothing about it. We therefore state the proposition as a conditional on contradictory provenance rather than as an existence claim about incompatible universes. We claim non-certifiability, not that the value of $\hat{y}$ can never coincide with a separately-valid answer. When outputs are natural-language or discrete, "combination" denotes any aggregation inheriting both provenances, not a vector convex combination. Three worked examples appear in Appendix M.

L Proof of Proposition 2: The Interior-Mixture Penalty

This appendix gives the full derivation behind Proposition 2 (§3), a self-contained argument for a K-component mixture of candidate expert distributions. It is an elementary risk argument, not a Bayesian-consistency result; we make no claim of posterior inconsistency or hypercompression. Epistemic incompatibility motivates the regime but is not a premise of Proposition 2; the theorem applies to any vertex-identifiable candidate family.

Setup. Let $u_1, \ldots, u_K$ be candidate universes — in our motivating setting, epistemically incompatible ones (Def. 1), though nothing below uses that property. Each induces a candidate predictive distribution $p_k(y \mid x)$. The mixture over a weight $w \in \Delta^{K-1}$ is

$$p_{\mathrm{mix}}(y \mid x; w) \;=\; \sum_{k=1}^{K} w_k \, p_k(y \mid x).$$

Let $P^*(y \mid x)$ be the true conditional. We work in the *vertex-identifiable* regime: for each x there is a correct universe $k^*(x)$ with $P^*(\cdot \mid x) = p_{k^*}(\cdot \mid x)$, and

$$p_{k^*}(\cdot \mid x) \;\notin\; \mathrm{conv}\{\, p_k(\cdot \mid x) : k \neq k^* \,\}, \tag{2}$$

i.e., the correct component is not reproducible by any mixture of the *other indexed components*, not merely distinct from each alternative. Note this is stated over the indexed set $\{k : k \neq k^*\}$ rather than as extremality of p_{k^*} in $\mathrm{conv}\{p_k\}_{k=1}^K$. The two agree when the components are pairwise distinct, but not in general: if $p_j = p_{k^*}$ for some $j \neq k^*$, then p_{k^*} is still an extreme *point* of the hull, yet (2) fails and the minimizing weight is not unique.

Why (2) **and not pairwise disagreement.** An earlier version of this argument assumed only that $p_k(\cdot \mid x) \neq p_{k^*}(\cdot \mid x)$ on a positive-measure set for every $k \neq k^*$, and concluded that $p_{\mathrm{mix}}(\cdot \mid x; w) = p_{k^*}(\cdot \mid x)$ forces $w = e_{k^*}$. That step does not follow. Pairwise disagreement constrains each component individually but says nothing about their mixtures: if $p_{k^*} = \frac{1}{2}p_2 + \frac{1}{2}p_3$ with $p_2 \neq p_{k^*} \neq p_3$, then $w = (0, \frac{1}{2}, \frac{1}{2}, 0, \ldots)$ reproduces P^* exactly and attains the same minimal risk, so the vertex is *not* the unique minimizer. Pairwise positive-measure disagreement is therefore not mixture-weight identifiability. Condition (2) is exactly what is required, and it is necessary as well as sufficient (see the converse below).

The vertex is the unique risk minimizer. Fix a strictly proper loss ℓ (e.g., log loss), for which the pointwise risk $R_x(q) = \mathbb{E}_{y \sim P^*(\cdot \mid x)}\, \ell(q, y)$ is uniquely minimized over predictive distributions q at $q = P^*(\cdot \mid x)$. For the mixture,

$$R(w) \;=\; \mathbb{E}_x \, \mathbb{E}_{y \sim P^*(\cdot \mid x)} \, \ell\big(p_{\mathrm{mix}}(\cdot \mid x; w),\, y\big).$$

At each x, R_x is minimized iff $p_{\mathrm{mix}}(\cdot \mid x; w) = P^*(\cdot \mid x) = p_{k^*}(\cdot \mid x)$. Suppose this holds with $\alpha := w_{k^*} < 1$. Rearranging $\alpha\, p_{k^*} + \sum_{k \neq k^*} w_k\, p_k = p_{k^*}$ gives

$$p_{k^*} \;=\; \sum_{k \neq k^*} \frac{w_k}{1-\alpha}\, p_k, \qquad \sum_{k \neq k^*} \frac{w_k}{1-\alpha} \;=\; \frac{1-\alpha}{1-\alpha} \;=\; 1,$$

so $p_{k^*} \in \mathrm{conv}\{p_k\}_{k \neq k^*}$, contradicting (2). Hence $\alpha = 1$ and $w = e_{k^*(x)}$ is the unique minimizer; every weight with $w_{k^*} < 1$ has strictly positive excess risk $R(w) - R(e_{k^*}) > 0$.

Converse. If (2) fails, the displayed identity runs the other way: writing $p_{k^*} = \sum_{k \neq k^*} \lambda_k p_k$ with $\lambda \in \Delta$ exhibits a *second* minimizing weight $w = (0, \lambda)$, so the vertex e_{k^*} is not unique and vertex-identifiability is the exact boundary of the claim rather than a convenient sufficient condition. We say "second minimizing weight" and not "interior weight" deliberately: if the failure is caused by a duplicated component, $p_j = p_{k^*}$ for some $j \neq k^*$, then $\lambda = e_j$ and the alternative optimum is itself a *vertex*. Uniqueness fails; degeneracy of the optimum does not follow.

Two familiar sufficient conditions imply (2): linear independence of $\{p_k(\cdot \mid x)\}_{k=1}^K$, and pairwise-disjoint supports. Disjoint supports are the situation that motivates us — a frequentist and a Bayesian answer to the same question are not merely different, and neither is a blend of the others — but we do *not* claim that epistemic incompatibility implies disjoint supports or (2). Incompatibility is a property of axiom sets; (2) is a property of predictive distributions, and we have no validated test that a real expert pool satisfies it (App. A, item iv). We state (2) as an independent premise because it is weaker than disjoint supports and is what the argument actually consumes.

Why soft gating must degenerate. An input-dependent soft gate $w(x)$ with temperature $\tau > 0$ and finite logits lies in the open simplex, so it never attains the vertex $e_{k^*(x)}$ exactly: it approaches the optimum only as its mass concentrates on $k^*(x)$, i.e., as $w_{k^*}(x) \to 1$. So in this regime the risk-optimal soft gate is precisely the one that has collapsed to a hard router; any residual mass on incompatible components is pure excess risk. This is the *interior-mixture penalty*.

Low temperature is neither necessary nor sufficient for that limit, and an earlier draft wrongly said the vertex is reached "only as $\tau \to 0$". Concentration is governed by the logit gap divided by the temperature. At fixed τ=1, logits $(M, 0, \ldots, 0)$ give $w_1 = e^M/(e^M{+}K{-}1) \to 1$ as $M \to \infty$, so the vertex is approached at constant temperature; conversely $\tau \to 0$ concentrates all mass on whichever component has the largest logit, which is the *wrong* vertex whenever $\arg\max_k z_k \neq k^*(x)$, and then drives excess risk up rather than to zero. The correct statement is the one above: the weight on the correct component must tend to one. Annealing τ is a way to get there only once the gate's ranking is already right, which is a property of the trained logits and not of the temperature. Nothing here establishes a training or architectural advantage for hard gates — consistent with §4.3, where we detect no hardness effect.

Contrast with the mixture-representable regime. Call a setting *mixture-representable* when $P^*(\cdot \mid x)$ lies in $\mathrm{conv}\{p_k(\cdot \mid x)\}$ and equals no single component, so only a blend reproduces it. (Interiority alone does not give the second half: a component that is itself a mixture of the others can sit in the interior and equal P^*, which is exactly the duplicate/redundant case that breaks uniqueness above. We therefore require non-coincidence explicitly rather than reading it off the position in the hull.) This is the setting Bayesian MoE targets (Bernardo & Smith, 2009; Li et al., 2025), and there an interior mixture strictly outperforms every vertex. Our proposition is the reverse regime, *vertex-identifiable*: P^* coincides with one component and is not reachable by blending the others, so averaging can only hurt. The two regimes are distinguished by where P^* sits relative to the hull, which is a statement about the candidate set — not about M-open/M-closed model-space assumptions, with which it should not be conflated (a vertex-identifiable set is technically M-closed, since the vertex recovers P^*; the content of the result is the penalty for *interior* weights, not misspecification of the family).

On gating entropy. Soft MoE gates on this task train to low entropy (Table 18). We record the observation and do not interpret it mechanistically: it concerns a classification gate rather than the output-space mixture this proposition governs, entropy does not track accuracy across the 20 matched runs (Spearman ρ=0.15, p=0.53; App. E), and the matched-gate comparison is null (§4.3). It corroborates nothing here.

Scope. This is a statement about the population risk of a fixed mixing rule in the vertex-identifiable regime. A soft gate whose mass concentrates on the correct component approaches the hard limit, so the proposition characterizes *when* explicit commitment is the safer design, not an impossibility for all soft schemes. It carries no empirical weight in this paper: the matched-capacity gate comparison (§4.3) is a *null* result, and it measures a classification gate rather than output-space mixing (see "Scope of the propositions"), so it neither confirms nor contradicts the proposition. Nor do we have a validated test for whether a given expert pool satisfies (2); establishing one is open. □

M Error Propagation: Empirical Demonstration

Proposition 1 concerns outputs with contradictory provenance; the demonstrations below illustrate how incompatible frameworks can generate such provenance when their conflicting assumptions are both active. We illustrate with three demonstrations using SciPy (Virtanen et al., 2020).

Demo 1: Coin Fairness (Freq vs. Bayes). Correct (frequentist): $z = 2.0$, $p = 0.0455$, reject H_0. Wrong (Bayesian): posterior Beta$(61, 41)$, $P(\theta > 0.5) = 0.977$. Mixed error: "4.6% probability the coin is fair"—conflating $P(\text{data} \mid H_0)$ with $P(H_0 \mid \text{data})$.

Demo 2: Parameter Estimation (Bayes vs. Freq). Correct (Bayesian conjugate): posterior $N(1.575, 0.25)$, CrI $[0.60, 2.56]$. Wrong (MLE): $\hat{\theta} = 2.10$, CI $[0.97, 3.23]$. Mixed error: "The CI gives 95% probability θ is in range"—Bayesian interpretation of frequentist object.

Demo 3: Atomic Stability (QM vs. Classical). Correct (quantum): ψ_{1s}, $E = -13.6$ eV, uncertainty principle prevents collapse. Wrong (classical): electron spirals into nucleus in $\sim 10^{-11}$s. Mixed error: "electron orbits at definite position *and* obeys $\Delta x \cdot \Delta p \geq \hbar/2$"—logical contradiction.

N Cloud / Generalist-LLM Evaluation Details

Version caveat. Every number in this appendix is on EpistemicBench **v4**. Of the 360 stratified items used throughout, 105 (29%) contain text that the v5 repair rewrites (App. V); over the full v4 standard test split the figure is $372/1{,}442 = 25.8\%$. The subset is therefore *not* byte-identical between versions, and these comparisons cannot be restated as v5 results. We retain them as a legacy contextual comparison. No claim in §4.1–4.3 depends on them.

N.1 Few-shot 7–8B LLMs and Mixture-of-Agents (v4 stratified-360)

The gap concentrates on `CROSS_FRAMEWORK` (LLMs 13–23% vs. ours 83%). `STAT_ILL_POSED` is the instructive column: 2024 7–8B models miss it ($\leq$ 13%), 2025 frontier models recognize it well (DeepSeek-v4-pro 86.7%, GPT-4.1 76.7%), and our router's 97% is in-distribution — dropping to 10.2% out-of-distribution, below TF-IDF. Read together with §4.2, where the router scores $0/30$ on certified ill-posed items, on genuinely novel ill-posed questions a frontier LLM likely matches or beats our router, and the 97% is a benchmark artifact rather than a capability. Additional shots *reduce* accuracy for the 2024 7–8B family (Liu et al., 2024). ($^\flat$MoA runs under the same 4-bit NF4 quantization as the few-shot setup, biasing MoA *conservatively*.)

N.2 v4 stratified-360 zero-shot (apples-to-apples)

*Qwen3-Instruct ships in a reasoning-mode ("thinking-token") configuration that emits a multi-line internal monologue before any final answer. With our zero-shot label-only prompt and $T{=}0$ greedy decoding the parser saw `UNKNOWN` on every output (0.0% accuracy). We treat this as a prompt-format mismatch rather than a model failure and exclude it from the comparison; an instruction-tuned variant of the same model family (Qwen2.5-7B) is included above. Re-prompting with the `/no_think` directive or post-hoc thinking-token stripping would likely recover a usable result and is left as future work. ‡ Router `STAT_ILL_POSED` accuracy is reported on the full v4 test split ($n_{\text{ILL_POSED}}{=}100$, single seed) to give a tight estimate; on the 360-sample subset ($n{=}30$ for this universe) the result is consistent within sampling error.

Table 8: LLM baselines on EpistemicBench v4 (stratified 360-sample subset, 30 per universe; same subset for all rows): 7–8B few-shot LLMs, 2025 zero-shot frontier LLMs, Mixture-of-Agents (Wang et al., 2024a), and our 465M fine-tuned router. Wilson 95% CIs in brackets. Our router's numbers here (single-best 95.56, mean 94.72±0.76) are on these 360 samples; the full-v4-test figures are the 95.42 deterministic checkpoint and the 94.51 ± 0.41 5-seed mean (n=1,442). Full 2026 micro-LLM sweep: App. F.

Model	Params	Shots/u	Accuracy	Latency (s)
2024 7–8B LLMs				
Qwen2.5-Instruct	7B	3	72.78 [67.9, 77.1]	0.46
Qwen2.5-Instruct	7B	8	66.94 [61.9, 71.6]	0.55
Llama-3.1-Instruct	8B	3	69.17 [64.2, 73.7]	2.54
Llama-3.1-Instruct	8B	8	65.83 [60.8, 70.6]	3.14
2025 frontier LLMs (zero-shot)				
DeepSeek-v4-pro	API	0	82.78 [78.5, 86.3]	5.7
GPT-4.1	API	0	75.83 [71.2, 80.0]	0.6
Soft combination of frontier LLMs				
MoA majority-vote (3 proposers)[♭]	∼22B	0	62.22 [57.1, 67.1]	1.7
MoA aggregator (Wang et al., 2024a)[♭]	∼29B	0	64.17 [59.1, 68.9]	2.6
MoA majority-vote (few-shot proposers)[♭]	∼22B	1	68.33 [63.4, 72.9]	2.0
MoA aggregator (few-shot proposers)[♭]	∼29B	1	65.28 [60.2, 70.0]	2.9
MoA majority-vote (few-shot proposers)[♭]	∼22B	2	74.17 [69.4, 78.4]	2.2
MoA aggregator (few-shot proposers)[♭]	∼29B	2	69.44 [64.5, 74.0]	3.1
Ours (FT, single best seed)	**465M**	—	**95.56 [92.9, 97.2]**	**0.023**
Ours (FT, 5-seed mean)	465M	—	94.72 ± 0.76	0.023
Ours (FT, ×5 ensemble)	465M	—	95.00 [92.2, 96.9]	0.113

Table 9: Zero-shot generalist LLMs on EpistemicBench v4 (stratified 360-sample subset, 30 per universe; same subset as Table 8). Decoding: T=0, label-only instruction. Wilson 95% CIs are reported in brackets (n=360 for overall, n=30 per universe; n=100 for the router's `STAT_ILL_POSED` column from the full v4 test). The overall-accuracy ranking Mixtral > Llama-3.1 ∼ Qwen2.5 has Llama and Qwen overlapping (49.6–59.8 vs 46.0–56.2), so we report the qualitative ranking as "Mixtral best, ∼34pp gap to our router" ($95.6 - 61.9 = 33.7$pp overall) rather than ranking the two smaller LLMs against each other.

Model	Params	Acc.	95% CI	ILL_POSED	Lat. (s)
Qwen2.5-Instruct (0-shot)	7B	51.1%	[46.0, 56.2]	13.3% [5.3, 29.7]	0.32
Llama-3.1-Instruct (0-shot)	8B	54.7%	[49.6, 59.8]	3.3% [0.6, 16.7]	3.24
Qwen3-Instruct (0-shot)*	8B	—	—	—	—
Mixtral-Instruct (0-shot)	46.7B	61.9%	[56.8, 66.8]	53.3% [36.1, 69.8]	18.12
Ours (FT, single best)	**465M**	**95.6%**	**[92.9, 97.2]**	**97.0%**[‡] **[91.5, 99.0]**	**0.023**
Ours (FT, ×5 ensemble)	465M	95.0%	[92.2, 96.9]	97.0%[‡] [91.5, 99.0]	0.113

N.3 Pilot n=109 cloud generalists (legacy summary)

The v4-stratified comparison in Table 9 is the primary apples-to-apples evidence; the earlier n=109 pilot comparison is summarized here in one paragraph (its full tables are superseded by v4 and omitted). On the pilot split our 465M router reached 97.25% at 16ms and ∼$0.0003/query, versus 87.96–96.33% for six cloud generalists (80B–1T parameters) at 1.4–12.4s and $0.003–0.058/query; only DeepSeek-v3.1 differed significantly from the router (p=0.010, McNemar's test (McNemar, 1947; Dietterich, 1998); each pilot universe has only ∼10 questions, so per-universe CIs are wide).

The corresponding latency–accuracy Pareto plot was removed for the same reason: it anchored on the pilot's wider accuracy range.

O Dataset Construction Protocol

Table 10: Universe definitions with boundary criteria. The left column is the label string as released; the body of the paper names these Frequentist, Bayesian, Mixed/Philosophical, Ill-posed, classical mechanics, Quantum and relativity.

Universe	Inclusion	Exclusion
STAT_FREQ	p-values, CIs, hypothesis tests	Priors, posteriors, credible intervals
STAT_BAYES	Prior specification, posterior computation	p-values, frequentist CI
STAT_MIXED	Compares frameworks or philosophical	Single-framework answer
STAT_ILL_POSED	Missing information, unanswerable	Answerable with assumptions
PHYS_CLASSICAL	Newtonian mechanics, $v \ll c$	$\hbar$, relativistic speeds
PHYS_QUANTUM	Wavefunctions, uncertainty	Classical solutions
PHYS_RELATIVITY	Lorentz, $v \approx c$	Non-relativistic

Annotation protocol (pilot v0, 685 samples). For the original 685-sample pilot dataset (which preceded v4 and is now used only as an architecture-comparison sanity check during development), two annotators independently labeled all 685 samples. Initial agreement: 94.2% (Cohen's $\kappa = 0.91$); disagreements (40 samples) resolved through discussion. EpistemicBench v4 (22,009 samples) uses a rule-based pipeline with deterministic source-to-universe mapping and per-universe spot-check audits ($\sim$5% noise reported in Limitations).

P Continual Learning: Extended Analysis

Replay is what controls forgetting in this setting, and the size of the remaining forgetting depends on which experiment you read. The three-seed sweep is the estimate we quote, because it is the only one with a spread: $75.9 \pm 5.9\%$ at no replay falling to $9.9 \pm 7.2\%$ at 10% (Table 12). The coarse single-run comparison (Table 11) puts naive fine-tuning and EWC at 86.84% and 75.00%, which agrees in direction, but reports 0.00% for the same 10% rehearsal condition that the sweep measures at 9.9%. We do not reconcile the two: they are separate runs, the coarse one is unreplicated, and its zero sits outside the sweep's seed range. Read $9.9 \pm 7.2\%$ as the effect and the coarse table as the direction only.

P.1 Replay Buffer Sensitivity (8 ratios × 3 seeds)

Table 11: Coarse continual-learning comparison (5→7 universe expansion).

Method	Old Acc	Overall	Forgetting
Naive Fine-tuning	11.84%	37.61%	86.84%
EWC (λ=1000)	23.68%	45.87%	75.00%
Rehearsal (10%)	**98.68%**	**97.25%**	**0.00%**

The exponential decay model $f(r) = 69.2 \cdot e^{-0.326\,r} + 7.9$ achieves $R^2 = 0.946$ over the fitted 0–10% window. Solving it for 10% forgetting gives $r_c = 10.7\%$, so we describe the threshold as *about* 11% rather than exactly 10%; the sweep's own 8% and 9% rows already sit at 11.0 and 11.4%, within one seed spread of the line. The fit also has a floor of 7.9%, so it does not predict forgetting reaching zero at any replay ratio — which is the respect in which Table 11's 0.00% and Table 13's $0.0\%/-2.22\%$ are not reproduced by the seeded sweep.

Table 12: Fine-grained replay sensitivity (mean ± std across 3 seeds). Forgetting decays exponentially, crossing 10% near 8–10% replay.

Replay	Old (%)	New (%)	Overall (%)	Forgetting (%)
0%	23.2±6.5	93.9±2.5	44.6±4.3	75.9±5.9
1%	33.8±5.5	97.0±2.5	52.9±4.4	65.4±5.4
2%	67.1±14.0	92.9±1.4	74.9±9.5	32.0±13.7
3%	57.2±3.3	93.9±0.0	68.3±2.3	41.5±3.3
7%	79.4±4.5	93.9±0.0	83.8±3.1	19.7±4.9
8%	88.2±7.5	96.0±2.9	90.5±4.4	11.0±7.2
9%	87.7±6.6	96.0±1.4	90.2±4.3	11.4±6.5
10%	88.8±7.2	93.9±3.0	90.4±4.1	9.9±7.2

Beyond 10% replay. The sweep was run out to 20% replay, but Table 12 and the fit above cover the 0–10% window only, and the higher ratios are not reported here: at three seeds per ratio the between-seed spread (±7pp at 10%) is larger than the differences between adjacent ratios, so individual points in that range are not separable. An earlier draft singled out one of them as dipping below its neighbors and described the dip as falling inside a ± 1.4pp band around the fit; the comparison was to a point outside the fitted window and the two figures do not in fact bracket each other, so we withdraw it. Nothing in the paper depends on the shape of the curve past 10%.

P.2 Expansion Order Robustness

At 20% replay the outcome does not depend on the order in which the new universes are added: all three orders retain at least 97.78% old-universe accuracy and none shows positive forgetting (Table 13). These are single runs at one replay ratio, and their $\approx 0\%$ forgetting is below the 7.9% floor of the seeded fit above, so we read them as evidence that order does not matter and not as a measurement of how much forgetting remains.

Table 13: Expansion order sensitivity (20% replay).

Order	Old Acc	Overall	Forgetting
Stats First	100.0%	100.0%	0.0%
Physics First	97.78%	98.68%	−2.22%
Mixed	100.0%	100.0%	0.0%

Q External Validation: MMLU Details

6 subcategories from MMLU (Hendrycks et al., 2021): high school statistics (216), high school physics (151), conceptual physics (235), college physics (102), electrical engineering (145), astronomy (152). Ground truth by subcategory mapping with keyword refinement. Router confidence on correct predictions is significantly higher than errors (0.951 vs. 0.876, $p < 0.001$).

R Worked Case Study: Double-Slit and the Behavioral Effect of Routing

To make the end-to-end conformance result of Section 4.5 concrete, we walk through one question across the three solver-conditioning conditions.

Question (true universe `PHYS_QUANTUM`): *"In a double-slit experiment with single electrons fired one at a time, how do we explain the build-up of an interference pattern on the detector screen, and what does this say about the electron's trajectory?"*

(M) Monolithic, no routing. A fixed Qwen2.5-7B-Instruct solver responds: *"Each electron travels through one of the two slits and the resulting intensity pattern on the screen builds up because of constructive and destructive interference of the wave-like distribution of many electrons over time."* Conformance score on `PHYS_QUANTUM` anchors: 0/4 (no *wavefunction*, no *superposition*, no *operator*, no *—psi⟩*). The answer reads as classical wave optics applied to many particles—a defensible-sounding but axiomatically wrong account, because under classical mechanics each electron has a definite trajectory.

(N) Name-only routing (prepended `[Framework: PHYS_QUANTUM]`). The same solver responds: *"This is a quantum mechanics problem. The electron exhibits wave-particle duality: as a wave it passes through both slits, and as a particle it lands on the screen."* Conformance: 1/4 (*wave-particle*). Closer in vocabulary, but no operators or superposition reasoning.

(A) Name + axioms (prepended `[Framework: PHYS_QUANTUM][Rules: Use quantum mechanics; operators, superposition, |ψ⟩; position/momentum noncommute]`). The solver now responds: *"Each electron is described by a wavefunction ψ that is in a superposition of 'passing through slit 1' and 'passing through slit 2'. The probability of detection at position x is $|\psi_1(x) + \psi_2(x)|^2$, the squared modulus of the coherent sum—this is the source of the interference. The electron does not have a definite trajectory in the classical sense; asking which slit it went through is operationally meaningless without a which-path measurement, which would itself collapse the superposition and destroy the pattern."* Conformance: 4/4. The answer now invokes wavefunction, superposition, and the noncommutativity of position and trajectory measurements—axioms that the monolithic answer denied.

What this example does and does not show. It shows that axiom-conditioning changes the generated explanation and its terminology in this instance: the (A) response invokes a wavefunction, a superposition and a which-path measurement where the (M) response did not. It does *not* show that the underlying reasoning mode changed. Conformance counts anchor terms, so a solver that adopts the vocabulary without the inference scores identically to one that reasons correctly, and in aggregate (Table 14) conditioning *lowers* final-answer accuracy. We select this case because the monolithic default is confidently wrong, which makes the contrast legible; it is an illustration, not evidence, and n=1.

S Adversarial Robustness, Uncertainty, Conformal Prediction, Scaling, Ablations: Tables

This appendix collects the tables for the experiments compressed in App. B and the Discussion.

S.1 End-to-End Solver Conditioning: Full Table

Under injection-blind scoring only the name-only condition improves on the monolithic baseline ($+36\%$) while name-plus-axioms falls 78% below it, and both conditioning arms end with lower final-answer accuracy than the monolithic 0.429 (Table 14).

S.2 Adversarial Robustness

On the v4 test split the fine-tuned router holds overall attack success rate to 1.53% against TF-IDF's 65.75%, and keyword injection is the only attack that still succeeds against it (Table 15).

S.3 Deep Ensemble Uncertainty Quantification

On this legacy v4 split the deep ensemble adds no accuracy over the best single seed (95.42% against 95.49%), and its mutual information discriminates errors from correct predictions (AUROC 0.938) but not boundary universes (AUROC 0.279; Table 16).

Table 14: End-to-end conditioning of a fixed solver (Qwen2.5-7B) by router-supplied universe information, 240 stratified questions (20 per universe). *Blind* = injection-blind conformance (anchors already present in that question's prompt are excluded from scoring). Raw conformance rises monotonically, but the (A) gain reverses under blind scoring; (N), which injects no axiom text, is the only condition with a genuine blind gain, and it trades away answer accuracy. Blind scoring is deliberately conservative under (A): removing *every* provided anchor also denies credit to a solver that correctly adopts the prompt's terms, so (A)'s collapsed blind score reflects that mechanical exclusion, not proof of parroting. Blind name-only gains concentrate on `MATH_NON_EUCLIDEAN` (+0.25), `CAUSAL_SCM` (+0.26) and `STAT_ILL_POSED`/`STAT_MIXED` (+0.11 each); with n=20 per universe (Wilson 95% CI ±0.21) these rankings are indicative only.

Condition	Raw conf.	Blind conf.	Δ blind vs (M)	Ans. acc.	Latency (s)
(M) Monolithic	0.236	0.153	—	0.429	1.92
(N) Name only	0.347	**0.208**	**+36%**	0.345	4.25
(A) Name + axioms	0.453	0.033	−78%	0.381	7.18

Table 15: Adversarial robustness on the v4 test split (Attack Success Rate ↓). The fine-tuned router is 43× more robust than TF-IDF (Wang et al., 2022). Attack-construction details are in App. B and the prompt templates appear with the rest of the reproducibility configs (Appendix T).

Attack Type	TF-IDF	Ours
Synonym Substitution	61.47%	**0.00%**
Keyword Injection	89.91%	**4.59%**
Mixed Language	45.87%	**0.00%**
Overall ASR	65.75%	**1.53%**

S.4 Conformal Prediction: Standard vs. Variance-Scaled Scores

Both variants are split conformal on the 5-seed ensemble's averaged probabilities $f(x)$. The standard nonconformity score is $s(x, y) = 1 - f_y(x)$. The variance-scaled score divides it by the ensemble's mutual information at that input, $s_{\text{vs}}(x, y) = (1 - f_y(x))/(\text{MI}(x) + \epsilon)$ with ϵ=10^{-6}, so an input the ensemble disagrees about is charged less for the same probability gap. In both cases $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the calibration scores and the prediction set is $\{y : s(x, y) \leq \hat{q}\}$, which can be empty.

S.5 Ablations and Calibration Details

Diagnostic-subset breakdown: EasyRoute (cross-domain, 680 samples, 96.6% acc), HardRoute (intra-domain incompatible pairs, 980 samples, 97.0%), Boundary (ambiguous/ill-posed, 320 samples, 98.1%), RealWorld (real open-source data only, 638 samples, 94.8%). **Hardest-pair analysis:** PHYS_QUANTUM vs PHYS_RELATIVITY 92.5%; PHYS_CLASSICAL vs PHYS_QUANTUM 92.9%; STAT_FREQ vs STAT_BAYES 99.5% (easiest). **Calibration:** ECE = 2.75% single, 2.17% ensemble; temperature scaling (Guo et al., 2017) with T^*=1.496 gives ECE 2.74% (marginal). **OOD detection:** confidence-thresholded ensemble outputs achieve AUROC = 0.968 (Hendrycks & Gimpel, 2016). **Fisher distance:** pairwise Fisher Information distances (Amari, 2016) yield a between/within separability ratio of 2.10×.

T Reproducibility

Classification head and pooling. The backbone returns last-hidden states $H \in \mathbb{R}^{T \times d}$ (d=1024). We mean-pool over non-padding tokens to obtain $h \in \mathbb{R}^d$, apply LayerNorm, dropout (0.1), and a single linear projection $W \in \mathbb{R}^{12 \times d}$ to produce class logits.

Optimization. Cross-entropy with label smoothing 0.05, AdamW (lr 2×10^{-4} for LoRA, 1×10^{-3} for the head), batch size 16, 5 epochs, linear warmup over 5% of steps, weight decay 0.01, FP32 throughout (FP16 collapses accuracy to 18.99% through gradient overflow). Only the LoRA

Table 16: Deep-ensemble uncertainty on EpistemicBench v4 (n=1,442), over the 5 *original production* seeds (distinct from the 5 independent retraining seeds used for the seed-controlled routing comparison in Table 21). MI = mutual information (epistemic uncertainty). The ensemble matches the best single seed in raw accuracy, and its mutual information *separates* errors from correct predictions (AUROC 0.938). That is discrimination, not calibration: the calibration evidence is the ECE below, and it is measured on this legacy split, not on the certified sets.

Metric	**Test (n=1,442)**
Single model (best seed)	95.49%
Individual mean ± std (5 seeds)	94.80±0.59%
Ensemble (avg softmax)	95.42%
MI ratio (wrong/correct)	**10.7**×
AUROC: MI → misclassification	**0.938**
AUROC: MI → boundary universe	0.279
Spearman(MI, confidence)	−0.997

Table 17: Standard split conformal vs. variance-scaled conformal. Calibration set n=99, test set n=109. *Mean set size is reported because larger sets buy coverage trivially*: variance scaling adds 0.01–0.07 classes per input and is not obtaining its coverage that way. Sizes below 1 reflect the empty sets in the same row. Coverage falls below the nominal $1-\alpha$ for $\alpha \geq 0.05$ under both scores, so we read this table as a comparison between the two, not as evidence that either attains its stated level on this small, distribution-shifted test set.

	Standard			**Variance-scaled**		
α	Coverage	Mean $\lvert C\rvert$	Empty	Coverage	Mean $\lvert C\rvert$	Empty
0.01	98.2%	1.02	0	99.1%	1.03	0
0.05	93.6%	0.94	7	96.3%	0.98	2
0.10	84.4%	0.84	17	89.9%	0.92	9
0.15	82.6%	0.83	19	86.2%	0.88	13
0.20	81.7%	0.82	20	81.7%	0.84	18

adapters and the 12-class head are trainable (1.6M parameters); the backbone is frozen. The deep ensemble uses seeds 42, 123, 456, 789, 1024; the seed-controlled comparisons of Table 21 use 5 independent retraining seeds.

Inference cost and software. Batch-1 routing latency is 26.5 ± 4.5 ms (median 25.5, p95 34.3; 100 runs) at 1,771 MB of allocated GPU memory on an RTX 3090, and throughput is flat at ≈ 39 queries/s from batch 1 to batch 32, so the router is latency-bound rather than throughput-bound. These were measured on an earlier checkpoint of the same architecture (`router_augmented_v1`), not on the v5 checkpoints of Table 1; we report them because the architecture and parameter count are identical, and we do not claim them as measurements of the v5 runs. The environment specifies lower bounds rather than pins: PyTorch ≥ 2.1, Transformers ≥ 4.36, PEFT ≥ 0.7, NumPy ≥ 1.24, SciPy ≥ 1.11, scikit-learn ≥ 1.3. Upstream corpora are taken at the Hugging Face revision current in August 2026; the canonical Hendrycks MATH copy has since been withdrawn (Ethics Statement), so that split is not re-downloadable at the recorded revision.

U Full Routing Comparison (v4-Unified)

The complete matched-capacity routing comparison summarized in §4.3 (all off-the-shelf, matched, Bayesian, and hard rows, with gating entropy and latency).

V Benchmark Artifact Audit and the v4→v5 Repair

We audit EpistemicBench for surface features that predict the universe without carrying epistemic content, over six families the label should be independent of: character/word length, per-character punctuation counts, non-ASCII and Greek codepoints, LaTeX markup, first- and last-k tokens, and a

Table 18: Routing on EpistemicBench v4 (n=1,442). The "(matched)" rows and Fair MoE use 3.7M trainable parameters — 2.3× the hard router's 1.6M, a budget *advantage* to the soft side — and report mean±std over 3 training seeds (paired significance: Table 21); gate H is the seed mean; Bayesian gates marginalize over K=8 passes (multi-seed analysis: Figure 2). "STE hard top-1" reimplements Molavi et al. (2026) at 3.7M. Latencies: measured batch-1 mean on an idle RTX 3090 (200 queries after warmup); *the ×5 ensemble latency is a 5× sequential-forward proxy (the original ensemble checkpoints are deprecated). ††Lexical baseline; drops to 82.2% on the hard split. "CF" = capacity-factor routing; "LB" = load-balancing auxiliary loss; "H" is gating entropy; "Fair MoE" = soft top-2 MoE at the matched 3.7M budget (the capacity-matched soft baseline). Legacy v2/v3 numbers: Appendix C.

Method	**Type**	**Acc.**	**Gate H**	**Lat.**
Random uniform	Floor	8.3%	–	–
Majority class	Floor	15.6%	–	–
Hash routing	Floor	8.3%	–	–
Embed. cosine sim.	Retrieval	91.7%	–	14.3ms
TF-IDF + LogReg††	Keyword	95.8%	–	0.1ms
Expert-Choice (CF)	Std. MoE	58.7%‡	0.246	–
Switch Transf. (top-1+LB)	Std. MoE	80.3%	0.025	–
Soft MoE (top-2)	Std. MoE	80.9%	0.000	38.2ms
Soft MoE (top-1)	Std. MoE	82.5%	0.015	–
Fair MoE (LoRA-matched)	Std. MoE	**93.50±0.56%**	**0.064**	–
Bayesian Gaussian†	Bayesian MoE	82.9%‡	–	–
Bayesian MC-Dropout†	Bayesian MoE	85.8%‡	–	–
Bayesian Temperature†	Bayesian MoE	84.0%	–	–
Bayes. Temp.-Stoch. (matched)	Bayesian MoE	93.57±0.42%	0.024	–
Bayes. Logit-Gauss. (matched)	Bayesian MoE	93.62±0.55%	0.054	–
Bayes. MC-Dropout (matched)	Bayesian MoE	93.64±0.58%	0.068	25.2ms
STE hard top-1 (MoR-LoRA-style)	Hard MoE	93.78±0.49%	0.089	–
Hard router (ours, 1.6M)	Epistemic	**94.51±0.41%**	–	**22.7ms**
Ensemble ×5 (avg. softmax)	Epistemic+UQ	≈95.0%	–	113.1ms*

content-blind skeleton in which every letter is mapped to `a`, every digit to `0`, and punctuation and whitespace are preserved. We also fit two unions: the four count-based families together, and all six together. (The high-purity repeated n-grams reported below are a separate diagnostic, not a fitted family.) Each family is fit on train and scored on test, so a family that is predictive *and* transfers is a stylistic correlate of the domain, while one that is predictive in train and shifts at test is the failure mode that produces an apparent worst-group collapse.

Table 19: Surface-only classifiers on the source-ID-disjoint test set (n=6,447; majority class `MATH_EUCLIDEAN` = 56.6%). The content-blind row cannot see a single content word.

Feature family	**Micro**	**Macro**	**Worst-grp**
length / Unicode / LaTeX	52.4–57.6%	7.8–11.9	0.0
punctuation counts	62.6%	24.5	0.0
affixes (first/last 3 tokens)	70.3%	41.6	6.3
shape (content-blind)	**76.1%**	42.6	6.0
the four count families, joined	62.3%	31.6	1.2
all six families, joined	72.6%	48.3	10.9
majority-class baseline	56.6%	8.3	0.0

What v5 repairs. Two artifacts are label-aligned in train and nearly absent in test, and both are removed uniformly across all twelve universes: deficiency-announcing parentheticals (90.6% of STAT_ILL_POSED train items vs. 0.24% of other universes' train items; 5.5% of its test items) and (Variant N) augmentation tags (40–87% of train items in six universes; 0% of test). The parenthetical pattern was lexical, and it missed four rarer variants phrased with "not defined"/"Undefined" (*"(The sample space is not defined.)"* 12, *"(Undefined concept in standard statistics.)"* 11, *"(Causal structure not defined; correlation ≠ causation.)"* 8, *"(X is not defined; no distribution specified.)"* 1). They remain in v5 on $32/340$ STAT_ILL_POSED train items (9.4%) and on no other item in any split, including test. They were found after v5 was frozen, are disclosed rather than patched, and are scheduled for the next release. Whitespace is normalized only inside items that were actually edited — a global normalization would rewrite ∼2.2k LaTeX-heavy MATH_EUCLIDEAN items and confound the ablation with a formatting change.

Table 20: **Artifact ablation** (v4 retained only for this comparison; five retraining seeds each, same test items). Both architectures move together. *The v4 and v5 columns are seed means ± sample SD; Δ is the mean paired difference from the hierarchical bootstrap, which recomputes the metric inside every resample.* For micro and macro, which are means, the two agree ($79.9 - 77.1 = +2.8$). Worst-group is a *minimum* over the twelve per-class recalls and so is not linear in the resample: its bootstrap mean sits below the difference of the displayed means ($+36.8$ vs. $56.9 - 18.3 = 38.6$; $+40.0$ vs. $57.8 - 16.4 = 41.4$). Both quantities are reported as computed — the displayed means are plug-in estimates and the Δ column is the bootstrap estimand the CIs belong to. The v4→v5 edit removes the deficiency parentheticals *and* the train-only augmentation tags, so every row is the effect of the combined edit, not of the marker in isolation.

Model	Metric	v4	v5	Δ (95% CI)
Ours	STAT_ILL_POSED	18.3 ± 14.3	66.1 ± 9.6	$+47.9$ $[+32.6, +60.8]$
	worst-group	18.3 ± 14.3	56.9 ± 7.0	$+36.8$ $[+22.5, +47.8]$
	macro	77.1 ± 1.5	79.9 ± 1.2	$+2.8$ $[+1.1, +4.7]$
	micro	88.07 ± 0.58	88.56 ± 0.09	$+0.5$ $[-0.2, +1.2]$
RoBERTa	STAT_ILL_POSED	16.4 ± 7.3	71.9 ± 7.1	$+55.4$ $[+46.2, +64.4]$
	worst-group	16.4 ± 7.3	57.8 ± 6.2	$+40.0$ $[+31.9, +47.6]$
	macro	77.0 ± 0.8	80.0 ± 1.8	$+3.0$ $[+1.2, +4.9]$

What v5 does not repair, and ships in KNOWN_LIMITATIONS.md instead. (i) Four universes — STAT_BAYES, STAT_MIXED_OR_PHILOSOPHY, STAT_ILL_POSED, CAUSAL_SCM — are 100% generator-produced on both sides of the split. (ii) Two test sets are near-degenerate: CAUSAL_SCM has 89% of its 74 items behind three 5-token prefixes, which is why every method scores ≈100% on it, and MATH_NON_EUCLIDEAN 51% of 151; the affected items carry template_degenerate: true so they can be excluded without deleting a universe. (iii) source_dataset alone predicts the universe for 78.7% of train items, and 57 of 89 source datasets map to exactly one universe. Table 22 reports every conclusion under subsets built from these frozen flags.

The STAT_MIXED trade-off is a taxonomy boundary. Repair costs STAT_MIXED_OR_PHILOSOPHY 11pp (ours; 13pp RoBERTa), and 105% of that loss is absorbed by STAT_ILL_POSED alone — no other class moves. The items every seed reassigns are *"Can a 95% confidence interval be interpreted as a probability?"* and *"Do p-values truly measure the evidence against a null hypothesis?"*, whose answers flip between frequentist and Bayesian readings, which is the STAT_ILL_POSED criterion; the items every seed keeps name both frameworks and ask for a comparison. The confusion is one-directional (0 of 128 STAT_ILL_POSED items move the other way). We therefore submit the 16 stable cases, with 16 length-matched STAT_MIXED controls and 16 length-matched STAT_ILL_POSED controls, to blind human adjudication; if labels change, we will release **v5.1** rather than edit v5 in place.

W Legacy v4 Specialist-vs-Routing-Variant Comparison

This is a legacy v4 table and it does *not* test the gating null. Its reference row is our 1.6M classification-head specialist, not a capacity-matched gate, so each contrast confounds gate with architecture and trainable budget — which is why the one CI here excluding zero belongs to the STE

Table 21: **Legacy v4 paired specialist-vs-routing-variant effects** (n=1,442). *This is not a matched-capacity comparison*: the reference row is our 1.6M classification-head specialist and each variant has 3.7M trainable parameters, so architecture and budget differ along with the gate. Pairing is seed-symmetric (seed-i vs. seed-i) over 5 retraining seeds; Δ is the mean paired accuracy difference (specialist − variant), with 95% CIs from a hierarchical cluster bootstrap over seeds *and* samples (B=10,000). The three-way capacity-matched gate comparison is on v5 (§4.3) and is null.

Variant (3.7M, matched)	Δ vs hard (pp)	95% CI	seed SD
Fair MoE (soft top-2)	$+0.75$	$[-0.03, +1.60]$	0.85
Bayes Temp-Stochastic	$+0.53$	$[-0.24, +1.26]$	0.77
STE hard top-1	$+0.67^*$	$[+0.12, +1.21]$	0.37

*CI excludes zero. For both *soft* gates the between-seed SD exceeds the effect: hard-vs-soft is inconclusive. STE is itself a *hard* MoE, so its $+0.67$pp is architecture, not hardness. Two further Bayesian gates at 3 seeds give $\approx +1.0$pp but are underpowered (App. C).

row, itself a *hard* gate. The gating null is established on v5 among the three 3.7M variants (Table 1, §4.3). We retain this table because it is the seed-symmetric v4 measurement of the *specialist-vs-variant* gap, on a different split (n=1,442) from the primary result; like every other such contrast in this paper it confounds architecture, objective and budget with the gate, and is not evidence that specialization is what produces the gap.

X Sensitivity of the Gating Null to Benchmark Artefacts

Every matched-gate contrast has a 95% interval containing zero in all four subsets, so the v5 gating null survives dropping the near-degenerate templates, the generator-produced items, and the universes that are generator-produced on both sides of the split (Table 22).

Table 22: **Sensitivity to benchmark artifacts.** Paired macro-accuracy effects (pp, 95% hierarchical bootstrap over seeds and samples). Subsets are defined *only* by metadata frozen in v5 — `template_degenerate`, `synthetic_source`, `universe` — so no model output enters their definition and they cannot be tuned. *artifact-filtered* drops items behind a top-three prefix in a near-degenerate class; *real-items* drops generator-produced items; *non-shared-generator* drops the four universes that are generator-produced on both sides of the split. Every matched-gate contrast includes zero in every subset.

Subset	n	soft − hard	Bayes − hard	Bayes − soft	ours − RoBERTa
v5-full	6447	-0.13 [-4.18, +3.67]	-0.17 [-1.75, +1.24]	-0.01 [-3.30, +2.91]	-0.10 [-1.80, +2.00]
artifact-filtered	6304	+0.12 [-3.44, +3.42]	+0.24 [-1.71, +1.99]	+0.08 [-2.78, +2.81]	+0.42 [-1.70, +3.28]
real-items	5547	+0.92 [-2.61, +3.89]	+0.72 [-1.90, +2.83]	-0.16 [-1.80, +1.41]	+0.74 [-1.55, +3.34]
non-shared-generator	5875	+0.08 [-3.15, +2.43]	+0.10 [-2.57, +2.43]	+0.07 [-1.89, +1.72]	+0.51 [-0.94, +2.04]

Y SCS-1: Semantic Counterfactual Stress Set

Status. SCS-1 is reported as a *machine-generated diagnostic*. Its protocol required a blind review by a qualified person with no prior exposure to this work before any confirmatory claim could be made from it; that review did not take place, and the fallback was recorded in the protocol before any item was generated, so that it could not be chosen after seeing results. No claim in the main text rests on SCS-1.

Construction. Eight generator families across four strata — framework framing (S1), withheld statistical input (S2), causal identifiability (S3) and assumption change (S4) — with no family reusing any of the five templates of §4.2. Every pair carries a machine-checkable certificate: for underdetermined members, two admissible completions that reproduce all stated quantities yet differ by at least a declared margin; for determinate members, a unique computed answer. A second implementation that does not import the generator recomputes every certificate by a deliberately different numerical route (Wilson bounds from the quadratic's roots, binomial p-values by explicit summation, Beta

quantiles by bisection, causal effects by rebuilding the joint distribution and checking it reproduces the reported conditionals). All 600 items verified; 0 failures, 0 exclusions.

Freezing and leakage. Inclusion rules, scoring script and item manifest were fixed before any model prediction; LOCKED is hashed in a manifest (`items_sha256 24d46773...`). Three leakage checks against v5 — exact normalized text, 60-character prefix, and 5-gram Jaccard ≥ 0.6 — return zero hits. A separate check confirms no item on either side carries the deficiency parenthetical of §4.1, any of its four residual variants, or the plain vocabulary of underdetermination, so a $0/n$ ill-posed recall here is a failure to generalize past a wording convention rather than a failure to read a cue.

Amendments. Three, each dated and each made before the run it affects. (1) Framing pairs ship in *cued* and *cue-ablated* phrasings after the first development output named its own framework; the ablated variant is primary. (2) Minimality is checked per pair kind, since framing twins must describe different frameworks and so cannot be near-duplicates. (3) Pair accuracy is scored on the outcome each stratum manipulates — framework for framing pairs, status for status pairs — because scoring status everywhere is satisfied on framing pairs by never predicting `STAT_ILL_POSED`, and every system then scored identically. The literal pre-registered reading is reported beside it.

Results. Ill-posed recall is $0/100$ for all eight systems (exact 95% upper bound 3.6%); determinate recall is $500/500$; same-status collapse is $300/300$. The same five marker-free checkpoints recall 56.2–82.0% of ill-posed items on the v5 test split (per seed: 64.1, 82.0, 65.6, 62.5, 56.2; mean 66.1 ± 9.6). Predictions are deterministic within (family $\times$ phrasing), so the effective number of independent observations is at most eight families $\times$ two phrasings rather than 300 pairs; the exact intervals over pairs are correspondingly too narrow, and under the pre-registered cluster bootstrap over generator families no system's pair accuracy is distinguishable from zero (e.g., seed 2, 16.7%, CI $[0.0, 33.3]$). Removing framework vocabulary drives four of five seeds to $0/100$ on framing pairs. The marker-trained checkpoint ties the best marker-free seeds, so the v5 repair removed a leak without buying semantic capability. Against the lexical baseline of §4.1 the specialist's pair-accuracy advantage is significant, but against a class-balanced variant of the same baseline it is not detected (p=0.45); we report both, since the rare-class weighting alone decides the comparison.

A generalist comparator on the same items. The reviewers of an earlier draft all raised the same objection: every system evaluated on the certified items is our own router or a lexical baseline, so a $0/n$ could be a property of items no system could decide. It is not. We ran `Qwen2.5-7B-Instruct` — the same solver as §4.5, untrained on this task — over both certified sets with one prompt fixed before the run (SHA-256 `57b40133...`), greedy decoding, 32 new tokens, and a strict parse that accepts only the two permitted JSON objects:

> Decide whether the question below, exactly as stated, has enough information and assumptions to have a unique answer. Return `"underdetermined"` if required information or an assumption is missing and different admissible choices could change the answer. Return `"determinate"` otherwise. Do not solve the question. Do not explain your decision. Reply with exactly one JSON object and no other text.

We ran two models under this protocol. `Qwen2.5-7B-Instruct` local and greedy (`temperature`=0, 32 new tokens), and `claude-sonnet-5` through the Anthropic API, which rejects the deprecated `temperature` parameter, so that run uses the provider default and we record the deviation rather than dropping it. Every one of the 520 calls returned a parseable status.

Neither generalist is a constant predictor: Sonnet answers "underdetermined" on 129 of 260 items and Qwen on 144, against the gold rate of 130, and Sonnet's $100/100$ specificity on SCS-1 and Qwen's $30/30$ on the twins rule out reading their recall as a bias toward one label. The first Sonnet pass used a 32-token output cap and returned no text for $95/260$ calls, with missingness concentrated in the determinate class (81 of 95); because that missingness depends on the gold label, neither scoring those rows wrong nor conditioning on the parsed subset estimates performance, so that pass is reported only as a parse failure and the numbers above come from the repeated run at a 4096-token cap with the identical prompt.

Two limits on what this licenses. The generalists are *told* the criterion and answer a binary question; our gate is given twelve labels and must infer that the status question is the one being asked. So this is not evidence that generalists route better — it is evidence that the certified items carry the

Table 23: Prompted generalists on the same certified items, against our router. Scope is matched: the SCS-1 rows are the 100 *status* pairs (strata S2 and S3, 200 items, gold balanced $100/100$), not all 300 SCS-1 pairs, and the twins rows are the 30 pairs of §4.2. *Pair* requires both members correct; *collapse* counts pairs given one status. Our router is the marker-free seed-1 checkpoint in both blocks. Its $100/100$ specificity beside $0/100$ recall is one fact, not two: it calls every item determinate.

System	Set	Ill-posed recall	Specificity	Pair	Collapse
`claude-sonnet-5`	SCS-1	**100/100**	**100/100**	**100/100**	**0/100**
	twins	26/30	27/30	24/30	5/30
`Qwen2.5-7B-Instruct`	SCS-1	100/100	75/100	75/100	25/100
	twins	19/30	30/30	19/30	11/30
ours (12-way gate)	SCS-1	0/100	100/100	0/100	100/100
	twins	0/30	12/30	0/30	25/30

information needed to decide them, which is the premise the $0/100$ and $0/30$ depend on and the one an earlier panel doubted. And these are two models on one prompt: we did not sweep prompts, and we report the first protocol we wrote down.