UNIVERSE ROUTING: WHY SELF-EVOLVING AGENTS NEED EPISTEMIC CONTROL

Zhaohui Geoffrey Wang*
USC Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089, USA
zwang000@usc.edu

ABSTRACT

A critical failure mode of current lifelong agents is not lack of knowledge, but the inability to decide *how* to reason. When an agent encounters "Is this coin fair?" it must recognize whether to invoke frequentist hypothesis testing or Bayesian posterior inference—frameworks that are **epistemologically incompatible**. Mixing them produces not minor errors, but structural failures that propagate across decision chains. We formalize this as the **universe routing problem**: classifying questions into mutually exclusive belief spaces before invoking specialized solvers. Our key findings challenge conventional assumptions: (1) hard routing to heterogeneous solvers matches soft MoE accuracy while being $7\times$ faster—because epistemically incompatible frameworks cannot be meaningfully averaged; (2) a 465M-parameter router achieves $2.3\times$ smaller generalization gap than keyword-matching baselines, indicating semantic rather than surface-level reasoning; (3) when expanding to new belief spaces, rehearsal-based continual learning achieves zero forgetting, outperforming EWC by 75pp—suggesting that modular epistemic architectures are fundamentally more amenable to lifelong learning than regularization-based approaches. These results point toward a broader architectural principle: reliable self-evolving agents may require an explicit **epistemic control layer** that governs reasoning framework selection.

1 INTRODUCTION

Consider an autonomous agent solving mathematical problems over extended deployment. It encounters two questions:

Q1: "A coin flipped 100 times shows 60 heads. Is it fair ($\alpha = 0.05$)?"
Q2: "Given uniform prior on bias θ, what is $P(\theta > 0.6 \mid 60 \text{ heads})$?"

Both involve coin flips. Both require probabilistic reasoning. Yet Q1 demands **frequentist statistics**—null hypothesis testing, p-values, and rejection regions—while Q2 demands **Bayesian statistics**—prior specification, likelihood computation, and posterior integration. These are not merely different methods for the same problem; they embody **incompatible epistemologies** with mutually exclusive axioms about the nature of probability itself (Jaynes, 2003; Berger, 1985).

For agents that autonomously chain reasoning steps, such epistemological confusion is not a minor error, but a **structural failure that propagates across decisions**. An agent that conflates frequentist and Bayesian reasoning will produce outputs like "the p-value is the probability the hypothesis is true"—a statement that is not wrong in degree, but wrong in kind. Worse, downstream reasoning steps that depend on this output inherit the corruption, as we demonstrate empirically in Appendix A.

This failure mode cannot be resolved by scaling. A larger language model with more parameters and training data may produce more fluent explanations, but fluency does not guarantee epistemic coherence. The problem is architectural: current agents lack an explicit mechanism for recognizing *which reasoning framework* a problem requires before attempting to solve it.

*ORCID: 0009-0006-1187-1903

Universe Routing. We formalize this challenge as classifying questions into K **belief space universes**—mutually exclusive reasoning frameworks, each with its own axioms and solvers:

- `STAT_FREQ`: Frequentist statistics (p-values, confidence intervals, hypothesis tests)
- `STAT_BAYES`: Bayesian statistics (priors, posteriors, credible intervals)
- `PHYS_CLASSICAL`, `PHYS_QUANTUM`, `PHYS_RELATIVITY`: Physics frameworks
- `STAT_MIXED`, `STAT_ILL_POSED`: Ambiguous or malformed questions

The term "universe" is deliberate: within each belief space, certain axioms hold and certain operations are valid. Crossing universe boundaries without explicit acknowledgment produces logical contradictions (Proposition 1).

Why Hard Routing? Traditional mixture-of-experts (MoE) architectures use **soft routing**—weighted combinations of expert outputs. This is appropriate when experts represent different skills applied to the same underlying reality. But for epistemically incompatible frameworks, soft routing is not merely inefficient; it is **semantically meaningless**. What does it mean to take 60% of a frequentist answer and 40% of a Bayesian answer? The result is not a compromise but an incoherence (Proposition 1).

Contributions. This work makes three claims:

1. **Epistemic routing is learnable**: Fine-tuned transformers (67M–465M) achieve 97–98% accuracy with 1.8–2.3× smaller generalization gaps than keyword baselines; a deep ensemble achieves 3.0× smaller gap. This consistency across architectures indicates semantic understanding rather than surface pattern matching.
2. **Hard routing is not a compromise**: For belief space universes, hard routing matches soft routing accuracy while being 7× faster—validating that these frameworks are geometrically separable in representation space.
3. **Modular architectures enable continual learning**: When expanding to new universes, rehearsal achieves zero forgetting while EWC fails (75% forgetting)—suggesting that explicit epistemic modularity is fundamentally more compatible with lifelong learning.

2 Related Work

LLM-based Agents. ReAct (Yao et al., 2023) interleaves reasoning and actions; Reflexion (Shinn et al., 2023) enables verbal self-correction. Recent surveys on self-evolving agents (Tao et al., 2024; Wang et al., 2024) identify **lifelong learning without catastrophic forgetting** as a central challenge (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017). Our work addresses a complementary problem: not how to retain knowledge, but how to select the appropriate reasoning framework for applying it.

Adaptive Routing. Adaptive-RAG (Jeong et al., 2024) routes queries to different retrieval strategies based on complexity. We extend routing from retrieval *strategy* to reasoning *framework*—a qualitatively different challenge because frameworks can be epistemologically incompatible. Lu et al. (2024) route between expert LLMs by query characteristics; we route to solvers with mutually exclusive axioms.

Mixture of Experts. Classical MoE (Shazeer et al., 2017; Fedus et al., 2022) and modern variants (Jiang et al., 2024) use soft or top-k routing to **homogeneous** experts that differ in specialization but share underlying assumptions. We demonstrate routing to **heterogeneous** solvers with mutually exclusive axioms—a setting where soft combination is not suboptimal but meaningless (Proposition 1).

Question Classification. Shao (2024) use DistilBERT to classify competition math problems into topic categories (algebra, geometry, etc.) within a single epistemological framework. We address a qualitatively different task: classifying across frameworks with incompatible axioms, where misclassification produces not wrong topic assignments but logically incoherent outputs (Proposition 1).

3 METHOD

3.1 PROBLEM FORMULATION

Definition 1 (Belief Space Universe)**.** *A belief space universe $u \in \mathcal{U}$ is a reasoning framework characterized by a tuple (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 $\exists\, a \in A_{u_i}, b \in A_{u_j}$ such that $b \Leftrightarrow \neg a$.*

Given question q, the routing problem is to predict $u^* = \arg\max_{u \in \mathcal{U}} P(u \mid q)$ where $\mathcal{U} = \{u_1, \ldots, u_K\}$ are mutually exclusive belief spaces. The router then invokes the corresponding solver $S_{u^*}(q)$.

Proposition 1 (Inconsistency of Framework Mixing)**.** *Let $u_i, u_j \in \mathcal{U}$ be epistemically incompatible. For question q with correct universe u_i, any convex combination $\hat{y} = \alpha S_{u_i}(q) + (1-\alpha) S_{u_j}(q)$ with $0 < \alpha < 1$ is semantically inconsistent: $\hat{y}$ does not belong to the validity domain V_u of any $u \in \mathcal{U}$.*

Proof sketch. By definition, $\exists\, a \in A_{u_i}, b \in A_{u_j}$ with $b \Leftrightarrow \neg a$. Output $y_i \in V_{u_i}$ requires axiom a; $y_j \in V_{u_j}$ requires $\neg a$. The combination $\hat{y}$ depends on both, so no consistent axiom set can justify it. We verify this empirically with three worked examples in Appendix A. □

3.2 DATASET CONSTRUCTION

We curate 685 samples across 7 universes using GPT-4 (OpenAI, 2023) generation with expert-designed constraints ensuring: (1) unambiguous ground truth labels, (2) diverse surface forms for the same underlying framework, and (3) balanced representation after augmentation. Split: 70% train (477), 15% validation (99), 15% test (109). An additional 56 out-of-distribution samples with novel phrasings form a separate unseen test set. Two annotators independently labeled all samples, achieving Cohen's $\kappa = 0.91$ (Landis & Koch, 1977) (details in Appendix E).

3.3 MODEL ARCHITECTURE

We fine-tune Qwen-1.5-0.5B (Bai et al., 2023) (465M parameters) with a classification head, and additionally evaluate BERT-base (Devlin et al., 2019) (110M), DistilBERT (Sanh et al., 2019) (67M), and RoBERTa-base (Liu et al., 2019) (125M). A critical implementation detail: **FP32 precision is essential**. FP16 training causes gradient overflow in the classification head, collapsing accuracy to 18.99% (near-random for 7 classes).

3.4 DESIGN RATIONALE: WHY HARD ROUTING IS A LOGICAL NECESSITY

Our choice of hard routing follows from Proposition 1:

(1) Universes are mutually exclusive by definition. A weighted average of incompatible frameworks is not partially correct but semantically incoherent.

(2) Soft routing degenerates to hard routing empirically. When frameworks are geometrically separable in representation space, the router assigns near-100% probability to one universe. Soft combination then reduces to hard selection with additional computational overhead.

(3) Hard routing enables modular expansion. New universes can be added by training only the router, without modifying existing solvers.

4 EXPERIMENTS

4.1 MAIN RESULTS: SEMANTIC UNDERSTANDING VS. KEYWORD MATCHING

Table 1 reveals a critical distinction. TF-IDF baselines achieve near-perfect test accuracy through keyword matching ("p-value" → frequentist), but accuracy drops by 26pp on held-out questions with novel phrasings. All four fine-tuned transformers—from 67M DistilBERT to 465M Qwen—achieve 1.8–2.3× smaller generalization gaps. This consistency across architectures indicates that **semantic understanding of epistemic boundaries**, rather than architecture or scale, drives the improvement.

Table 1: Generalization comparison (accuracy in %). Gap = test − unseen. Fine-tuned transformers achieve 1.8–2.3× smaller gaps than keyword baselines; ensemble achieves 3.0×.

Method	Params	Test Acc	Unseen Acc	Gap
Random Baseline	–	21.10%	14.29%	+6.81%
Logistic Reg. + TF-IDF	–	97.25%	71.43%	+25.82%
SVM + TF-IDF	–	98.17%	71.43%	+26.74%
BERT-base	110M	97.25%	82.14%	+15.11%
DistilBERT	67M	98.17%	83.93%	+14.24%
RoBERTa-base	125M	97.25%	85.71%	+11.54%
Qwen-1.5-0.5B	465M	97.25%	83.93%	+13.32%
Qwen ens. (×5)	465M	**98.17%**	**89.29%**	**+8.88%**

Table 2: Routing strategy comparison. Equal accuracy with 7× speedup validates that belief spaces are geometrically distinct—soft combination provides no benefit.

Routing Strategy	Accuracy	Inference Time
Soft (MoE-style weighted)	97.25%	38.2ms
Hard (argmax selection)	**97.25%**	**5.5ms**

A deep ensemble (Lakshminarayanan et al., 2017) of 5 Qwen models further reduces the gap to 8.88% (3.0× smaller than keyword baselines).

4.2 Hard vs. Soft Routing: Validating Epistemic Separability

Table 2 validates our architectural claim: soft routing provides **zero accuracy benefit** over hard routing while incurring 7× latency cost. This is consistent with Proposition 1: belief spaces are geometrically separable, so the router assigns near-deterministic probabilities, causing weighted combination to degenerate to selection.

4.3 Robustness: Semantic Understanding Resists Adversarial Manipulation

Table 3 demonstrates that semantic understanding is not merely more generalizable but more **robust**. Keyword injection (adding "consider the prior" to frequentist questions) fools TF-IDF baselines 90% of the time but our model only 4.6%. This robustness is critical for deployed agents that may encounter adversarial or confusingly-worded inputs.

4.4 Comparison with Large-Scale Cloud Models

We evaluate six large cloud models (80B–1T parameters) in zero-shot mode on our 109-sample test set. Our 465M router achieves 97.25% accuracy at 16ms latency, running 88–775× faster than all models tested. Only one model (DeepSeek-v3.1, 671B) differs significantly ($p = 0.010$, McNemar's test (McNemar, 1947)); the remaining five are not statistically distinguishable at $n = 109$. All cloud models struggle on `STAT_ILL_POSED` (64.7–94.1% vs. our 100%), suggesting that detecting ill-posed questions benefits from explicit boundary training. Full results in Appendix B.

4.5 External Validation on MMLU

To address concerns about generalization beyond synthetic training data, we evaluate on **1,001 real questions** from 6 MMLU subcategories (Hendrycks et al., 2021). A router trained on 477 synthetic questions outperforms TF-IDF by +10.6pp (56.8% vs. 46.2%). Accuracy improves monotonically with confidence: at ≥0.99 confidence, it reaches 70.7% on 617 samples. The gap from internal evaluation reflects coarse proxy labels and domain shift from synthetic to real phrasing. Details in Appendix C.

Table 3: Adversarial robustness (Attack Success Rate ↓). Keyword-based methods are trivially fooled; semantic understanding provides 43× better robustness.

Attack Type	TF-IDF Baseline	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%**

Table 4: Continual learning: expanding from 5 to 7 universes. Rehearsal achieves zero forgetting; EWC fails despite careful tuning.

Method	Old Universes	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%**

4.6 Continual Learning: Modularity Enables Expansion Without Forgetting

Table 4 reveals a striking asymmetry. EWC—a principled regularization approach—reduces forgetting only marginally (87% → 75%). Rehearsal with just 10% replay (29 samples) achieves **zero forgetting**. This suggests that **modular epistemic architectures are fundamentally more compatible with continual learning** than monolithic alternatives. The router's task—assigning questions to discrete universes—naturally decomposes into separable subproblems that rehearsal preserves; EWC's diagonal Fisher approximation cannot capture this structure.

Expansion order robustness: Performance is stable regardless of whether statistics or physics universes are learned first (0–2% variation), indicating robustness to curriculum effects (Appendix F).

5 Discussion

Epistemic confusion is structural, not a knowledge gap. Three observations support this. First, hard routing matches soft routing accuracy (Section 4.2), consistent with well-separated belief spaces. Second, semantic understanding provides adversarial robustness that keyword matching cannot (Section 4.3). Third, modular architecture enables continual learning that regularization methods cannot achieve (Section 4.6). These are not incremental improvements but qualitative capabilities that emerge from architectural choices.

Fine-tuning vs. scale. The consistency across four architectures (67M–465M, Table 1) and the competitive performance against zero-shot cloud models up to 1T parameters (Section 4.4) suggest that the advantage of fine-tuning for epistemic routing does not reduce to model capacity. The key ingredient is explicit boundary supervision, not scale.

Toward epistemic control as an architectural component. Just as modern agents have explicit memory systems and tool-use interfaces, reliable self-evolving agents may require an explicit layer that governs reasoning framework selection. Universe routing is one instantiation of this principle.

Limitations. Our dataset (685 samples) covers only 7 universes in mathematical and physical domains. Extension to broader reasoning frameworks (legal, ethical, causal) remains unexplored. Single-label hard routing cannot handle genuinely multi-step tasks that require crossing framework boundaries within a single problem—an important direction for future work. The small test set ($n = 109$) limits statistical power for cloud model comparisons. We evaluate routing accuracy, not end-to-end task performance with downstream solvers, though Appendix A demonstrates that framework confusion produces qualitatively wrong outputs.

6 CONCLUSION

We have argued that **adaptive reasoning framework selection is not an optimization problem, but a prerequisite for reliable self-evolving agents**. Proposition 1 formalizes the conditions under which mixing incompatible frameworks produces incoherent outputs. Our experiments support three claims: (1) epistemic routing is learnable with semantic rather than surface-level understanding across multiple architectures; (2) hard routing to incompatible frameworks is not a compromise but a logical necessity; (3) modular epistemic architectures are fundamentally more amenable to continual learning than monolithic alternatives.

These results suggest a direction for agent architecture: explicit epistemic control as a first-class component, governing not just *what* the agent knows but *how* it reasons. Universe routing is a first step toward this goal.

ACKNOWLEDGMENTS

The author thanks the anonymous LLA workshop reviewers for their constructive feedback that improved this paper.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985. doi: 10.1007/978-1-4757-4286-2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

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

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 7036–7050, 2024.

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6467–6476, 2017.

Keming Lu, Hongyi Yuan, Runji Lin, et al. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. doi: 10.18653/V1/2024.NAACL-LONG.109.

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

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Yourui Shao. An accurate classification and recommendation method of competitive math problems. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*, pp. 97–103, 2024. doi: 10.1109/BigDataService62917.2024.00021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*, 2017.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhengwei Tao, Ting-En Cheng, Jing Gao, Jiaxuan Liu, Zhiqiang Wang, Shan Jia, et al. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*, 2024.

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.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A ERROR PROPAGATION: EMPIRICAL DEMONSTRATION

Proposition 1 characterizes the conditions under which mixing incompatible solvers produces semantically inconsistent outputs. We verify this with three concrete demonstrations, computing actual numerical results under correct and incorrect frameworks using SciPy (Virtanen et al., 2020).

Demo 1: Coin Fairness (Freq vs. Bayes). Question: "A coin flipped 100 times shows 60 heads. Is it fair ($\alpha = 0.05$)?" The *correct* framework (frequentist) yields: $z = 2.0$, $p = 0.0455 < 0.05$, reject H_0. The *wrong* framework (Bayesian with Beta$(1, 1)$ prior) yields: posterior Beta$(61, 41)$, $P(\theta > 0.5 \mid \text{data}) = 0.977$. The *mixed* error: "There is a 4.6% probability the coin is fair"—conflating $P(\text{data} \mid H_0)$ with $P(H_0 \mid \text{data})$, a statement wrong in both frameworks simultaneously.

Demo 2: Parameter Estimation (Bayes vs. Freq). Question: "Given prior $\theta \sim N(0, 1)$ and observations $x = \{2.1, 1.9, 2.3\}$, find the posterior." Correct (Bayesian): posterior $N(1.575, 0.25)$, CrI $[0.60, 2.56]$. Wrong (frequentist MLE): $\hat{\theta} = 2.10$, CI $[0.97, 3.23]$. The mixed error: "The CI $[0.97, 3.23]$ gives 95% probability that θ lies in this range"—a Bayesian interpretation of a frequentist object.

Demo 3: Atomic Stability (Quantum vs. Classical). Question: "Why is the hydrogen atom stable?" Correct (quantum): stationary state ψ_{1s} with $E = -13.6$ eV; uncertainty principle prevents collapse. Wrong (classical): accelerating electron radiates continuously, spiraling into the nucleus in $\sim 10^{-11}$s. The mixed error: "The electron orbits at definite position *and* obeys $\Delta x \cdot \Delta p \geq \hbar/2$"—a logical contradiction.

In all three cases, the mixed output is not merely a worse approximation; it is a statement that is *wrong in both frameworks simultaneously*, instantiating Proposition 1.

B CLOUD MODEL EVALUATION DETAILS

We evaluate six cloud models via API: Qwen3-Next (80B), GPT-OSS (120B), Cogito-2.1 (671B), DeepSeek-v3.1 (671B), GLM-4.7 (696B), and Kimi-K2.5 (1T).

Table 5: Cloud model comparison (109-sample test set). †McNemar's test with continuity correction.

Model	Params	Acc.	Latency	Speedup	$p^\dagger$
Qwen3-Next	80B	96.33%	3,640ms	228×	1.000
GPT-OSS	120B	91.74%	1,986ms	124×	0.077
Cogito-2.1	671B	94.44%	1,413ms	88×	0.289
DeepSeek-v3.1	671B	87.96%	4,090ms	256×	**0.010**
GLM-4.7	696B	95.28%	12,392ms	775×	0.131
Kimi-K2.5	1T	94.50%	9,875ms	617×	0.371
Ours	**465M**	**97.25%**	**16ms**	**1×**	—

C EXTERNAL VALIDATION: MMLU DETAILS

We use 6 subcategories from MMLU (Hendrycks et al., 2021) totaling 1,001 test questions. Ground truth universes are assigned by subcategory mapping with keyword refinement. Router accuracy varies by alignment with our taxonomy: high school physics (82.1%), high school statistics (65.7%), college physics (61.8%), conceptual physics (60.4%), electrical engineering (60.0%), astronomy (7.2%). Astronomy's low accuracy reflects taxonomy mismatch (questions about black holes labeled PHYS_CLASSICAL; PHYS_RELATIVITY is arguably more appropriate). The router's mean confidence on correct predictions is significantly higher than on errors (0.951 vs. 0.876, $p < 0.001$), indicating informative uncertainty under distribution shift.

D Design Choices and Justifications

D.1 Why FP32 Training Is Essential

FP16 training consistently collapses to 18.99% accuracy regardless of learning rate, batch size, or warmup schedule. FP32 achieves 97.25%. Root cause: gradient overflow in the classification head during early training. Unlike language modeling where token-level losses are averaged over thousands of positions, classification concentrates gradient magnitude in a single output. We verified this across 5 random seeds—FP16 fails deterministically.

D.2 Why GPT-4 Augmentation Does Not Constitute Data Leakage

We do not ask GPT-4 to answer questions but to *generate* questions given universe constraints. Expert-designed structural requirements force novel combinations rather than memorized examples. All 200 generated samples were manually reviewed; 20 were rejected. The unseen test set was written after model training with novel phrasings. Without augmentation, rare classes suffered severe underfitting (73.39% $\rightarrow$ 97.25% improvement).

E Dataset Construction Protocol

Each universe is defined by necessary and sufficient conditions:

Table 6: Universe definitions with inclusion/exclusion criteria.

Universe	Inclusion Criteria	Exclusion Criteria
STAT_FREQ	Requires p-values, confidence intervals, hypothesis tests	Mentions priors, posteriors, credible intervals
STAT_BAYES	Requires prior specification, posterior computation	Asks for p-values or frequentist CI
STAT_MIXED	Explicitly compares frameworks or philosophical	Has clear single-framework answer
STAT_ILL_POSED	Missing information makes question unanswerable	Can be answered with assumptions
PHYS_CLASSICAL	Newtonian mechanics, thermodynamics	Involves $\hbar$, relativistic speeds
PHYS_QUANTUM	Wavefunctions, uncertainty principle	Can be solved classically
PHYS_RELATIVITY	Lorentz transformations, $v \approx c$	Non-relativistic regime

Two annotators independently labeled all 685 samples. Initial agreement: 94.2% (Cohen's $\kappa = 0.91$). Disagreements (40 samples) resolved through discussion; 12 relabeled, 3 removed.

F Continual Learning: Extended Analysis

F.1 Replay Buffer Sensitivity

Table 7: Sharp threshold in replay requirements.

Replay %	Samples	Old After	New Acc	Forgetting
0% (Naive)	0	11.84%	96.97%	86.84%
5%	14	59.21%	93.94%	39.47%
10%	**29**	**98.68%**	93.94%	**0.00%**
20%	59	93.42%	93.94%	5.26%

A sharp threshold exists between 5% and 10% replay. This suggests a critical mass of replay samples is required to maintain universe decision boundaries.

Table 8: Expansion order sensitivity (all with 20% replay).

Order	Phase 1 → Phase 2	Old Acc	Overall	Forgetting
Stats First	STAT_*, PHYS_CL → PHYS_Q, PHYS_R	100.0%	100.0%	0.0%
Physics First	PHYS_* → STAT_F, STAT_B	97.78%	98.68%	-2.22%
Mixed	STAT_F, PHYS_CL, STAT_B → PHYS_Q, PHYS_R	100.0%	100.0%	0.0%

F.2 Expansion Order Robustness

All orderings achieve <3% forgetting variation, confirming robustness to curriculum effects. Negative forgetting in "Physics First" indicates beneficial transfer.

G Error Analysis

Only 3 of 109 test samples (2.75%) are misclassified. All occur at genuine epistemic boundaries:

Error 1: Double-slit experiment classified as PHYS_CLASSICAL instead of PHYS_QUANTUM (confidence: 67%). Defensible—classical wave optics also explains interference.

Error 2: Factory defects classified as STAT_FREQ instead of STAT_BAYES (confidence: 73%). Question mixes frequentist terminology with Bayesian methods.

Error 3: "Meaning of statistical significance" classified as STAT_FREQ instead of STAT_MIXED (confidence: 81%). Model focuses on subject matter rather than meta-level question.

All error cases show lower confidence (67–81%) than correct predictions (mean 94%), suggesting calibrated uncertainty.

H Reproducibility

Parameter	Value
Base model	Qwen-1.5-0.5B
Precision	FP32 (required)
Optimizer	AdamW, lr=5×10^{-5}
Batch size	8, Epochs: 3
GPU	RTX 3090 (24GB)
Training time	4 minutes (single run)
Total for reproduction	∼20 GPU-hours

Code, dataset, and model checkpoints will be released upon publication.