ICML 2026

Models Under SCOPE

Scalable & Controllable Routing via Pre-hoc Reasoning

> Don't just pick a model — **predict how it will behave first.**

Qi Cao · Shuhao Zhang · Ruizhe Zhang · Ruiyi Zhang · Peijia Qin · Pengtao Xie

Routing saves cost — but today's routers are rigid

Most routers treat routing as closed-set classification: pick one name from a fixed list.

Send easy queries to cheap models, hard ones to strong models. Two limitations break this in practice:

Can't handle new models

A router trained on a fixed set doesn't know what to do with a freshly released model. Adding one means collecting new data and retraining the whole system.

No flexible control

Routers make a hard choice ("Model A") with no estimate of how much better or more expensive it is — so users can't trade accuracy for cost on the fly.

SCOPE: from selection to estimation

Two fundamental shifts that make routing scalable and controllable.

~~*Memorize model names*~~

→ Read behavioral fingerprints

Decide by how a model actually answered similar questions — so SCOPE can evaluate any model, even unseen ones, with no retraining.

~~*Output "pick Model A"*~~

→ Predict correctness + cost

Estimate if a model will be right and how many tokens it costs, then maximize a budget-aware utility — turning routing into a tunable decision.

The SCOPE framework — three stages

Construct fingerprints → predict outcomes → decide under budget.

STAGE A

Fingerprint Construction

Retrieve how the target model behaved on the most semantically similar anchor questions (from a 250-query anchor set).

STAGE B

Performance Prediction

A reasoning estimator predicts correctness $\hat{y}$ and token cost ℓ, trained with hindsight-distillation SFT then GRPO.

STAGE C

Decision & Calibration

A budget-aware utility, anchored to real historical performance, selects the optimal model per query.

Adapting to a new model = one forward pass over 250 anchors. No gradient updates.

Behavioral fingerprints, not model names

A model is described by what it does, so any model can be scored the same way.

1. Fix a compact anchor set of 250 representative queries spanning STEM → Humanities.

2. Record each model's ground-truth correctness and token cost on those anchors — its "fingerprint".

3. For a new query, retrieve the top-K most similar anchors and feed that behavioral slice to the estimator.

Why it matters

Training-free scalability.
New models are added by reading their fingerprint — no retraining, no labels, no gradient steps.

Predictions are grounded.
Estimates come from observed behavior on similar problems, not guessing from a name.

Training a reasoning estimator

Two stages on a Qwen3-4B backbone: a stable start, then sharpened accuracy.

STAGE 1 · SFT

Hindsight Distillation

A teacher sees the query and the realized outcome (y, ℓ), then writes a concise rationale justifying it. The model learns short, stable chains-of-thought.

STAGE 2 · RL

GRPO Alignment

A gated, format-checked reward combines a correctness term with an adaptive token-error tolerance — driving precise correctness and cost estimates.

77.0%

SCOPE prediction accuracy

75.1%

without chain-of-thought

59.4%

few-shot base model

One knob: trade accuracy against cost

A preference coefficient α turns routing into a controllable optimization.

Accuracy priority (high α)

Beat strong baselines: +25.7% accuracy over the unseen Claude-Sonnet-4.5 — while cutting its cost by 74%.

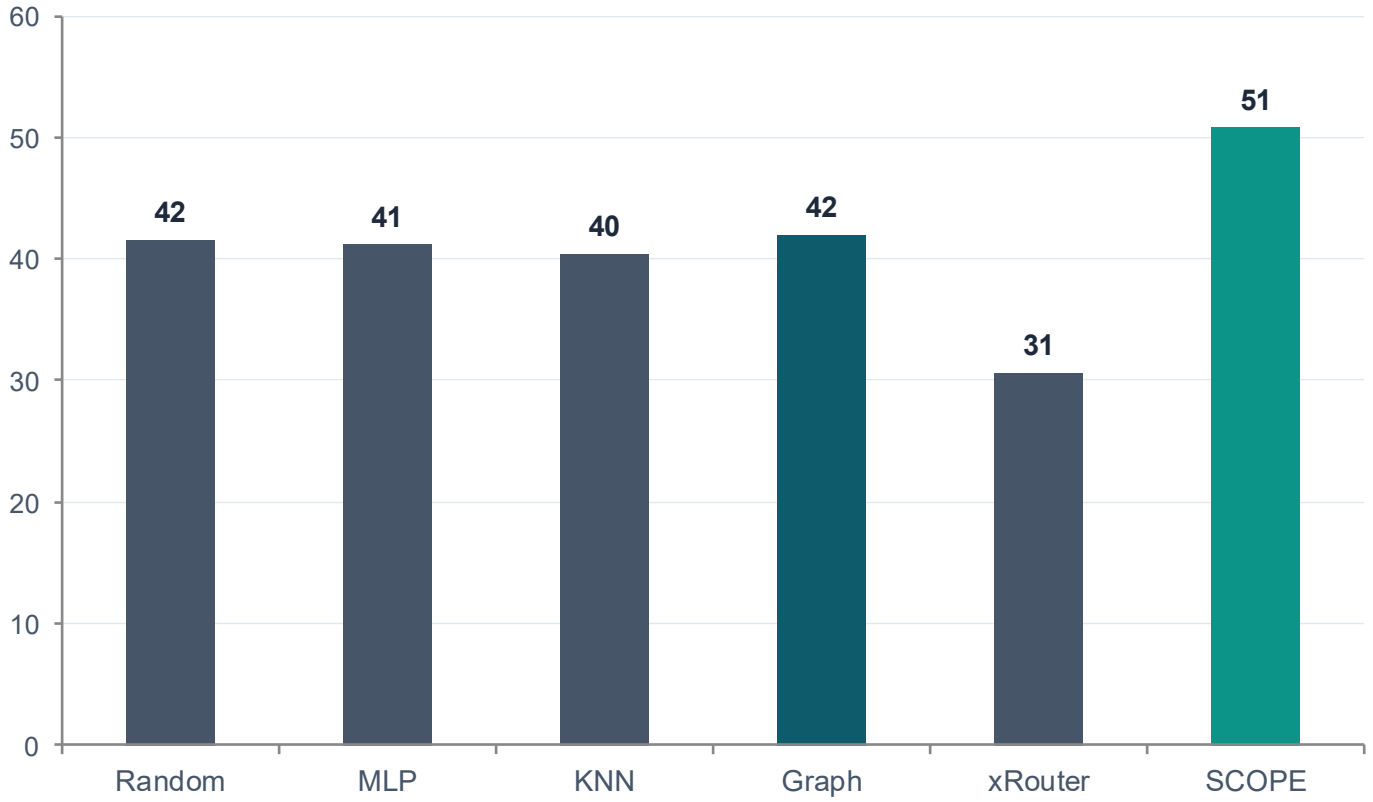
Cost priority (low α)

Stay competitive on accuracy while slashing inference cost by up to 95.1%.

Better trade-offs than any single model

Across seen (Test) and unseen (OOD) settings, SCOPE beats router baselines.

Average accuracy on the OOD set (unseen models)



Headline results

+24.3%

accuracy over Qwen3-235B (seen)

+25.7%

accuracy over Claude-Sonnet-4.5 (unseen)

−95.1%

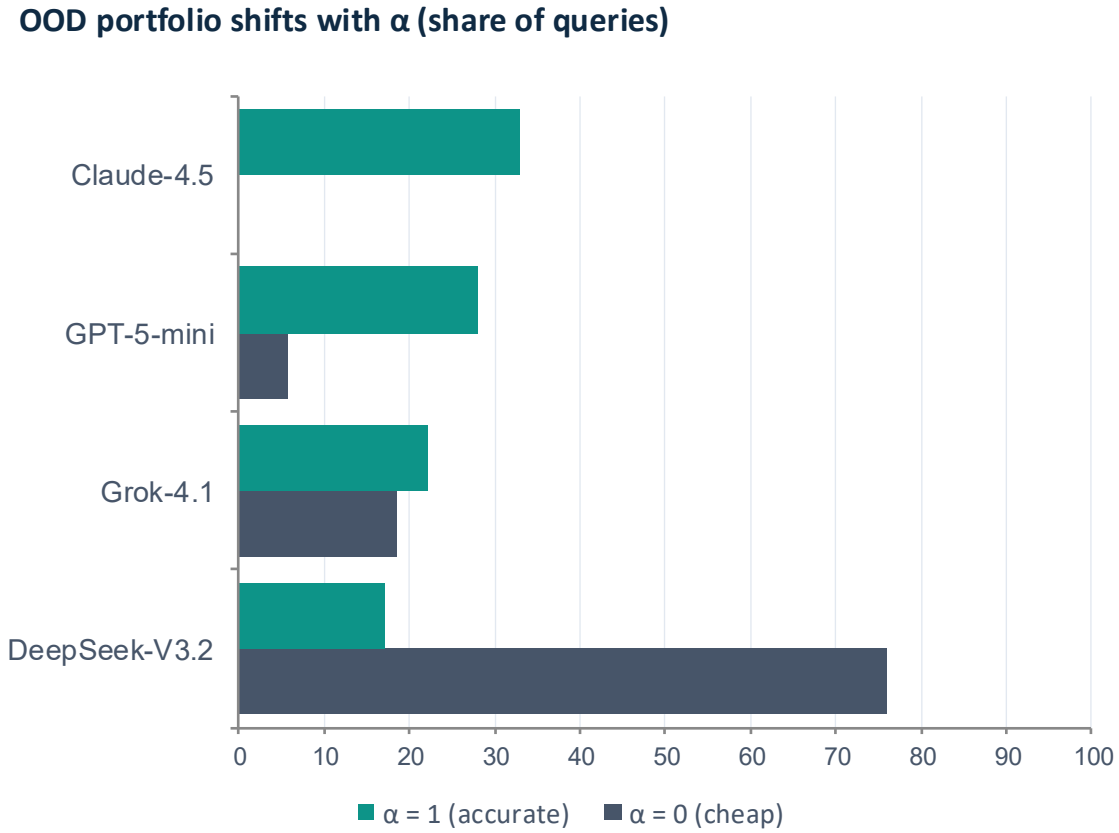
inference cost when efficiency is priority

It generalizes where classifiers fail

On unseen models, retrained baselines collapse to near-random — SCOPE adapts with no retraining.

Re-trained supervised routers **drop to near random** on unseen architectures due to distribution shift; xRouter's fixed pool can't include them at all.

By retrieving past behavior of similar models, SCOPE routes effectively and even **identifies powerful new models** like Claude-Sonnet-4.5 from their fingerprints alone.



Cheap to run, cheap to adapt

Pre-hoc prediction replaces running every model; anchor inference replaces retraining.

−90.8%

tokens vs. test-time scaling

1.8k vs 19.2k tokens per query with a 7-model pool — and the gap widens as the pool grows.

238.7

avg. prediction tokens

Hindsight distillation shrinks the reasoning trace from 2,355 → 239 tokens per prediction.

38×

less compute to adapt

Adapting to a new domain costs 9.0×10^{16} FLOPs (anchor inference) vs 3.4×10^{18} for retraining a router.

Takeaways

- Routing need not be a fixed choice — predict each model's correctness & cost first.
- Behavioral fingerprints generalize to unseen models with no retraining.
- One knob α gives users direct, per-query control over accuracy vs. cost.
- A practical form of pre-hoc test-time scaling: better trade-offs than any single model.

Released

Full pipeline on GitHub

Weights + SCOPE-60K & 250 on HuggingFace

Thank you!

Models Under SCOPE · ICML 2026

Project Page: https://sullivan07043.github.io/SCOPE/