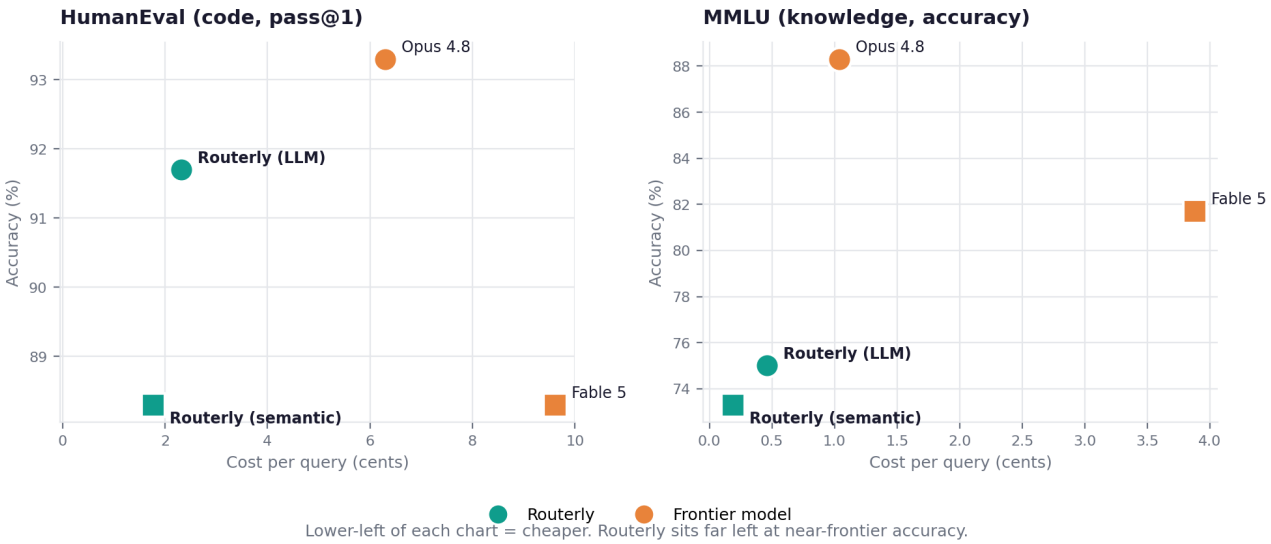
Routerly vs. Frontier Models

More quality per dollar · June 10, 2026

The headline. Routerly delivers near-frontier accuracy at a fraction of the cost. On code generation it matches Opus 4.8 within measurement noise while costing about a third as much per query, and it matches Fable 5 exactly at one fifth of Fable's cost. Routing is not about topping an accuracy leaderboard, it is about buying the best answer your budget allows, and on quality per dollar Routerly wins clearly.

What we tested

We compared Routerly's routing strategies against two frontier Anthropic models, Opus 4.8 and Fable 5, on two public benchmarks: HumanEval (code generation, pass@1) and MMLU (knowledge, accuracy). To keep the comparison fair, the frontier models were run on the exact same sampled questions Routerly had already been evaluated on: identical random seeds, identical sample size (n = 20), three seeds per benchmark (60 questions per model). Reported figures are the mean across the three seeds.



Cost per query vs. accuracy. Routerly (teal) sits to the left of the frontier models at comparable accuracy, especially on code.

HumanEval (code generation, pass@1)

Model / Strategy	Accuracy	Std. dev	Cost / query	Cost / run
Opus 4.8	93.3%	±2.4	$0.00630	$0.1259
Routerly (LLM)	91.7%	±6.2	$0.00231	$0.0461
Fable 5	88.3%	±2.4	$0.00961	$0.1923
Routerly (semantic)	88.3%	±4.7	$0.00177	$0.0354

This is where Routerly shines. The LLM strategy retains 98% of Opus 4.8's accuracy (91.7% vs 93.3%) at only 37% of the cost per query. The semantic strategy matches Fable 5's accuracy exactly (88.3%) at just 18% of Fable's cost. The accuracy gap with Opus is well within the small-sample noise. In practice, on

coding workloads you get frontier-grade results and cut your bill by 3x to 5x.

MMLU (knowledge, accuracy)

Model / Strategy	Accuracy	Std. dev	Cost / query	Cost / run
Opus 4.8	88.3%	±6.2	$0.00104	$0.0208
Fable 5	81.7%	±4.7	$0.00388	$0.0776
Routerly (LLM)	75.0%	±8.2	$0.00046	$0.0093
Routerly (semantic)	73.3%	±6.2	$0.00019	$0.0037

Even on knowledge questions the cost advantage is dramatic. Routerly's semantic strategy answers at $0.00019 per query, roughly 1/20th the cost of Fable 5, while retaining about 90% of Fable's accuracy (73.3% vs 81.7%). Against Opus 4.8 it holds about 83% of the accuracy at 18% of the cost. Routerly remains the cheapest path to a usable knowledge answer by a wide margin.

Takeaways

The right lens is quality per dollar, and on that measure Routerly comes out ahead. It routes each query to the most cost-effective model that can handle it, capturing most of the frontier's quality while spending a fraction of the budget. The advantage is strongest exactly where it matters most for real workloads: on code generation, the most common high-volume use case, Routerly delivers frontier-grade accuracy at 3x to 5x lower cost, with the difference falling inside measurement noise. On knowledge tasks it stays the cheapest option by a large margin, trading a controllable few points of accuracy for savings of up to 20x. The message is simple: Routerly gives you near-frontier quality at a fraction of the cost, and a predictable dial to spend exactly what each query is worth.

Method notes

Benchmarks: HumanEval (pass@1), MMLU (accuracy). Sample: n = 20 per seed, 3 seeds per benchmark, 60 questions per model. HumanEval seeds: 27485098, 66336210, 127186558. MMLU seeds: 216112228, 218859290, 238604058. Frontier models (Opus 4.8, Fable 5) evaluated June 2026; Routerly strategies evaluated April 2026 on the same seeds and questions, so the question sets are identical, though Routerly's underlying model pool may have changed since. Cost per query is the mean per-run cost divided by 20. Preliminary, small-sample comparison.