

LM4PLAN WORKSHOP · ICML 2026

AgentRouter: Heterogeneous Model Routing for Cost-Optimal Multi-Step Agentic Workflows

Rudrendu Kumar Paul · Sourav Nandy

Boston University University of Texas at Austin

Enterprise agents route every call to a frontier model — and waste 60–80% of inference spend doing it

THE HIDDEN WASTE IN AGENTIC PIPELINES

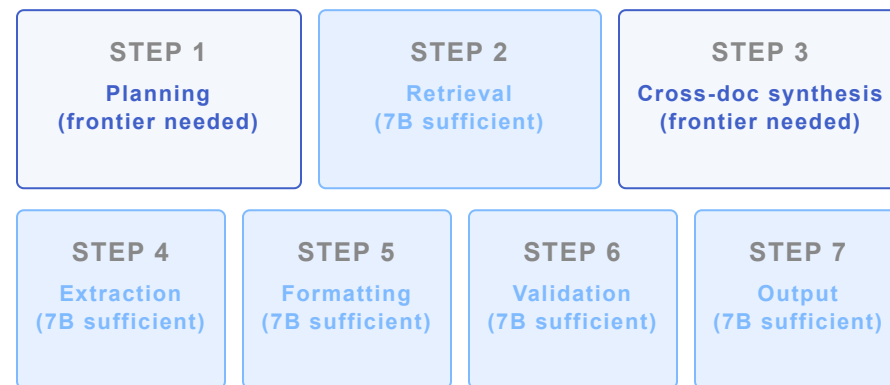
Existing routing solutions optimize single-turn query assignment but miss a property unique to agentic workflows: **subtask complexity varies widely within a single trajectory**.

In a 7-step document analysis workflow, only 2 steps need frontier-class reasoning. The other 5 produce identical quality with a 7B model.

Routing all 7 steps to GPT-4-class: \$0.42

Selective routing with AgentRouter: **\$0.18** — no measurable quality loss

STEP-LEVEL COMPLEXITY IN A 7-STEP WORKFLOW



☐ Frontier required ☐ 7B model sufficient

AgentRouter: a 12M-parameter classifier that assigns model tier per trajectory step in <5ms

SYSTEM DESIGN

AgentRouter formalizes step-level model routing as a **sequential assignment problem over agent trajectories**. The classifier operates at the boundary of each trajectory step — before any LLM call is dispatched.

12M

classifier parameters

<5msrouting overhead per
step (A100)**50K**annotated trajectory
steps in training set

Training spans planning, coding, research, and data analysis trajectories.

4 MODEL TIERS (ORDERED BY COST & CAPABILITY)

Tier	Profile	Example use
T1 Frontier	Highest capability	Multi-doc synthesis, planning
T2 Mid-high	Strong reasoning	Code generation, analysis
T3 Efficient	Balanced	Summarization, classification
T4 Minimal	7B-class, fast	Retrieval, formatting, extraction

5 FEATURES — NO LLM NEEDED AT ROUTING TIME

- Task type (f_type): plan / reason / retrieve / generate / format / verify — classified by a 2M-parameter BERT tagger
- Reasoning depth (f_depth): [0, 1] score from conditional-clause count and multi-hop indicators
- Tool-use requirements (f_tool): count of tool calls specified or implied in the step

Single-turn routers fail on trajectories — corrupted intermediate steps degrade every step that follows

THE GAP IN ROUTELLM AND FRUGALGPT

RouteLLM and FrugalGPT treat each LLM call as independent. That assumption holds for single-turn QA. It breaks in agentic workflows, where each step receives the output of the previous step as its input context.

A cheap model producing a subtly wrong intermediate result — a misattributed clause in a contract, a missed entity in an extraction — forces the frontier model at the next step to reason over **corrupted context**. The error propagates forward.

Trajectory-aware routing accounts for **inter-step quality dependencies**. AgentRouter models downstream impact at the moment of routing — before the call is dispatched.

CASCADING QUALITY LOSS — SINGLE-TURN VS. TRAJECTORY ROUTING

SINGLE-TURN ROUTER (ROUTELLM / FRUGALGPT)

Step N: cheap model → subtly wrong output
Step N+1: frontier model receives corrupted context
Step N+2: compound error in synthesis

End-to-end quality degraded — cost still paid

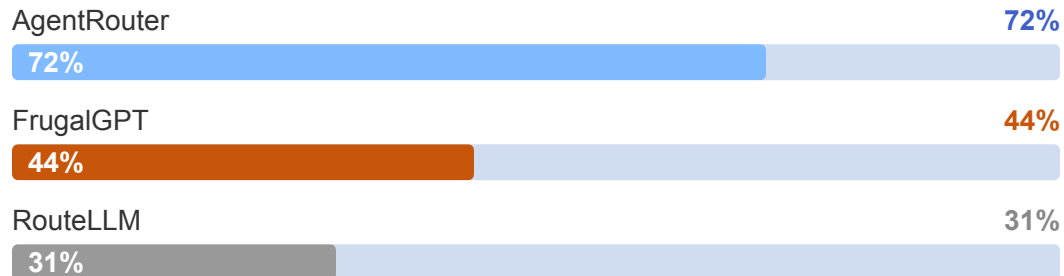
AGENTROUTER — TRAJECTORY-AWARE

Step N: router models downstream deps → correct tier assigned
Step N+1: clean context passed forward

72% cost reduction, <3% quality loss

72% cost reduction vs. frontier-only baselines — outperforming RouteLLM (31.4%) and FrugalGPT (44.1%)

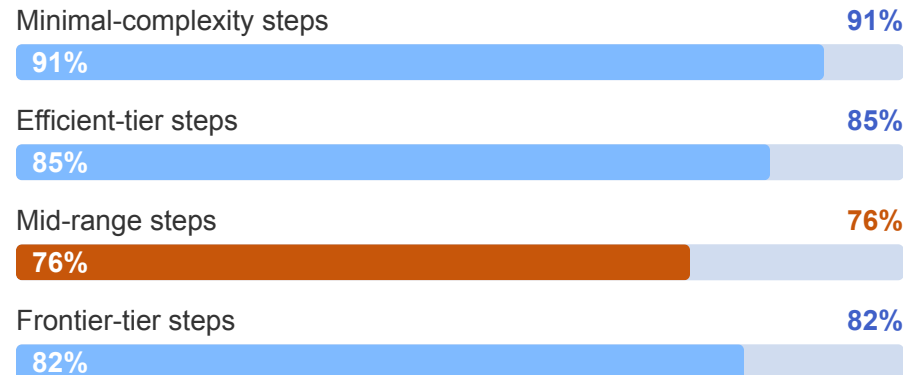
COST REDUCTION VS. FRONTIER-ONLY BASELINE



RouteLLM and FrugalGPT achieve lower reduction because their single-turn training signal misses trajectory-level quality dependencies.

AgentRouter: **<3% degradation** in end-to-end task completion vs. frontier-only.

PER-STEP ROUTING ACCURACY BY TIER



Mid-range-tier accuracy is lowest (76%) because the capability boundary between efficient and frontier is most ambiguous. The asymmetric loss biases the classifier toward over-routing to Tier 1 for borderline mid-range steps.

Match model to step complexity — the minority of steps that need a frontier model determine end-to-end quality

CONTRIBUTIONS

- Formalization of **step-level model routing** as a sequential assignment problem over agent trajectories
- AgentRouter: **12M-parameter classifier**, 5 lightweight features extractable at routing time, <5ms overhead per step on A100
- **4-tier model taxonomy** ordered by capability and cost, with per-tier routing accuracy reported
- **72% cost reduction** vs. frontier-only baseline with <3% end-to-end task completion loss
- Outperforms RouteLLM (31.4%) and FrugalGPT (44.1%) — baseline adaptations that treat each call as independent

KEY FINDING

55–70% of agent trajectory steps are satisfiable at sub-frontier capability. The minority that demand frontier reasoning — planning, cross-document synthesis, complex multi-step inference — determine end-to-end quality.

TAKEAWAY

Match model to step complexity. Step complexity determines when frontier reasoning is warranted; task-level assignment wastes it. A frontier model called at the right moment is decisive — called at every moment, it is just expensive.