Router-R1: Teaching LLMs Multi-Round Routing and Aggregation via Reinforcement Learning

Haozhen Zhang, Tao Feng, Jiaxuan You

University of Illinois at Urbana-Champaign

Presenter: Haozhen Zhang

Background – ***What Is The Problem?***

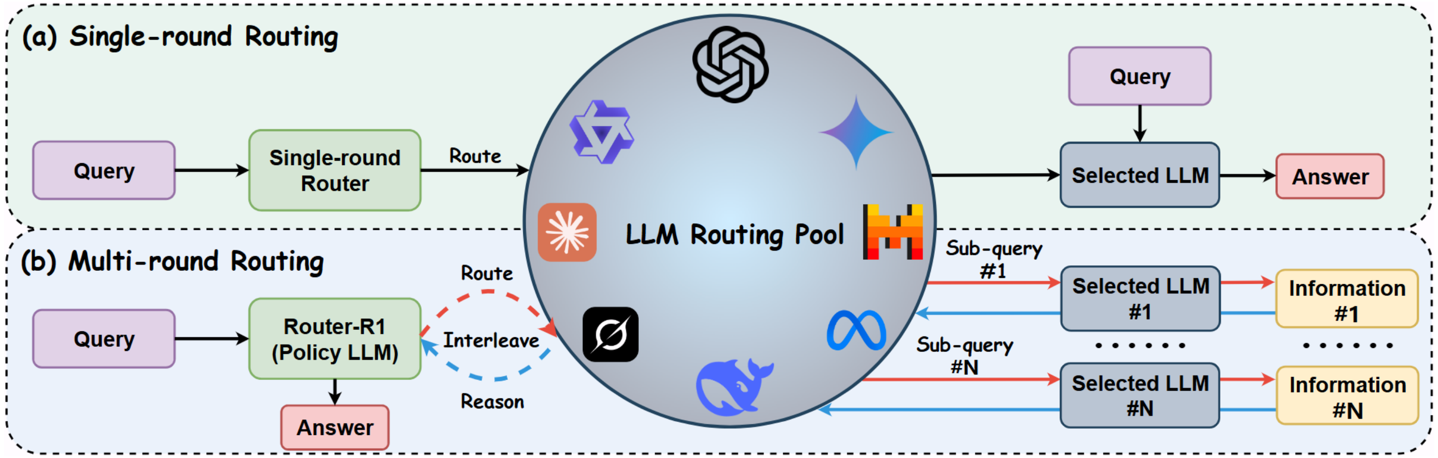
- The LLM ecosystem has rapidly expanded — **dozens of models now coexist**.
- **User queries vary drastically** in difficulty and domain.
- Sending *all* queries to a single large model causes:
  - ***Suboptimal performance***
    - Each LLM has its own specialized domain or reasoning strength.
    - A “one-size-fits-all” approach ignores this diversity.
  - ***Resource inefficiency***
    - Complex models are overused for trivial queries.
    - Simple tasks could be handled by smaller, cheaper models.

Background – ***Related Work***

- **Early single-round routers:** RouterDC, GraphRouter…
- **Core idea:** select *one* LLM per query
- Limitations:
  - Lack of multi-step coordination — cannot combine complementary LLMs.
  - Static and one-shot decisions → no feedback or reasoning loop.

Method – ***Overview***

- Prior routers (RouterDC, GraphRouter) → effective but **single-round**.
- They overlook a key fact:
  - ➢ Complex reasoning often requires **multiple coordinated model calls**.
- **Goal:** orchestrate multi-model collaboration through *multi-round routing + aggregation*.
- Inspired by **DeepSeek** and **Search-R1**, Router-R1 introduces the **first multi-round router**.



Method – ***Overview***

- **Router Initialization:** a capable LLM itself (e.g., Qwen or LLaMA).
- Candidate LLM descriptions are injected into the prompt as *cold-start knowledge*.
- Through training, Router-R1 learns the **strengths & weaknesses** of each model.

Answer the given question. Every time you receive new information, you must first conduct reasoning inside `<think>` and `</think>`.
After reasoning, if you find you lack some knowledge, you can call a specialized LLM by writing a query inside `<search> Candidate LLM: Query </search>`.
Before each LLM call, you must explicitly reason inside `<think>` and `</think>` about "why external information is needed" and "which LLM from the list is most suitable for answering your query," based on the brief model descriptions provided below.
When you call an LLM, the response will be returned between `<info>` and `</info>`. You are encouraged to explore and utilize different LLMs multiple times to better understand their respective strengths and weaknesses, as well as gather more comprehensive information.
Description of LLM Candidates: {candidates_intro}
If you find that no further external knowledge is needed, you can directly provide your final answer inside `<answer>` and `</answer>`, without additional explanation or illustration.
Question: {question}

Method – ***Reward Curation***

- **Optimization:** Proximal Policy Optimization (PPO)
- **Reward components:**
  - **Format Reward** – encourages correct <think>/<route> syntax.
  - **Final Outcome Reward** – based on answer correctness (EM).
  - **Cost Reward** – inverse to (API price × #output tokens).

$$\mathbf{R}_{\text{format}} = \begin{cases} -1, & \text{if the format is incorrect} \\ 0, & \text{if the format is correct} \end{cases}$$

$$\mathbf{R}_{\text{outcome}} = \mathbf{EM}(y_a, g_t),$$

$$\mathbf{R}_{\text{cost}} \propto -m(P_{\text{LLM}}) \cdot T_{\text{out}},$$

- **Overall Reward:** $r_\phi(x, y) = \mathbf{R}_{\text{format}} + (1-\alpha)\mathbf{R}_{\text{outcome}} + \alpha\mathbf{R}_{\text{cost}}$

  where α controls the performance–cost trade-off.

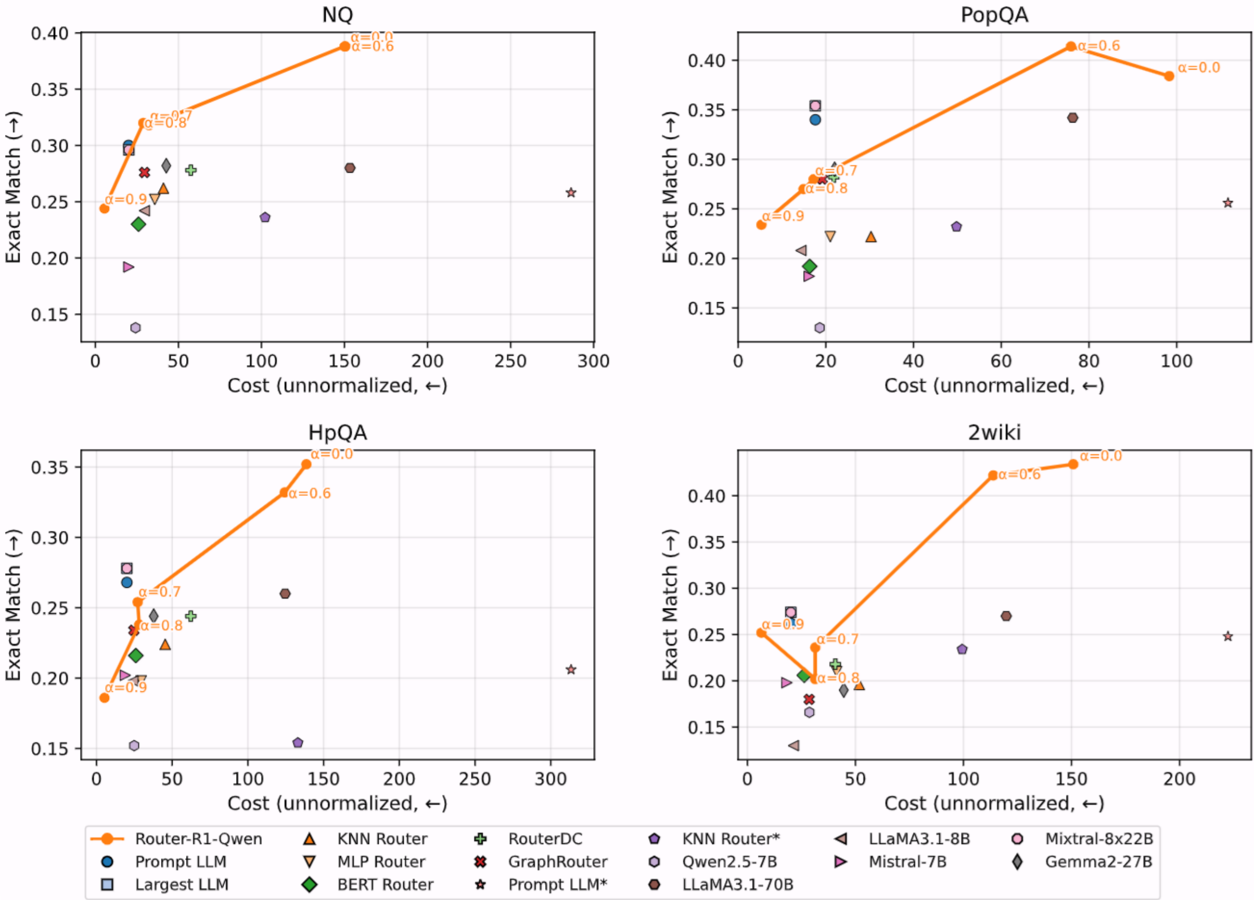
- Enables the router to **balance accuracy and efficiency** during RL training.

Experiments

Methods	General QA			Multi-Hop QA				
	NQ†	TriviaQA	PopQA	HpQA†	2wiki	Musique	Bamb	Avg.
Qwen2.5-3B-Instruct								
Direct	0.092	0.260	0.122	0.140	0.266	0.026	0.040	0.135
CoT	0.126	0.358	0.160	0.168	0.208	0.046	0.224	0.184
SFT	0.212	0.400	0.160	0.198	0.256	0.052	0.112	0.199
RAG	0.298	0.540	0.366	0.216	0.146	0.078	0.224	0.267
Search-R1	0.328	0.510	0.324	0.236	0.278	0.090	0.272	0.291
Prompt LLM	0.300	0.580	0.340	0.268	0.262	0.108	0.448	0.329
Largest LLM	0.296	0.578	0.354	0.278	0.274	0.104	0.480	0.338
KNN Router	0.262	0.528	0.222	0.224	0.196	0.066	0.360	0.265
MLP Router	0.252	0.460	0.222	0.198	0.210	0.072	0.360	0.253
BERT Router	0.230	0.516	0.192	0.216	0.206	0.058	0.312	0.247
RouterDC	0.278	0.592	0.282	0.244	0.218	0.080	0.504	0.314
GraphRouter	0.276	0.586	0.280	0.234	0.180	0.076	0.448	0.297
Prompt LLM*	0.258	0.500	0.256	0.206	0.248	0.078	0.472	0.288
KNN Router*	0.236	0.478	0.232	0.154	0.234	0.072	0.384	0.256
Router-R1-Qwen	**0.388**	**0.706**	**0.384**	**0.352**	**0.434**	**0.138**	**0.512**	**0.416**
Llama-3.2-3B-Instruct								
Direct	0.202	0.328	0.176	0.144	0.134	0.018	0.048	0.150
CoT	0.256	0.468	0.182	0.172	0.168	0.040	0.272	0.223
SFT	0.076	0.098	0.084	0.100	0.224	0.026	0.016	0.089
RAG	0.308	0.478	0.356	0.162	0.084	0.038	0.176	0.229
Search-R1	0.372	0.578	0.360	0.282	0.226	0.084	0.272	0.311
Prompt LLM	0.304	0.638	0.374	0.248	0.198	**0.132**	**0.528**	0.346
Largest LLM	0.344	0.616	0.394	0.258	0.242	0.122	0.472	0.350
KNN Router	0.292	0.572	0.254	0.210	0.182	0.078	0.376	0.281
MLP Router	0.282	0.506	0.248	0.178	0.188	0.064	0.360	0.261
BERT Router	0.256	0.560	0.222	0.210	0.188	0.066	0.296	0.257
RouterDC	0.310	0.614	0.298	0.250	0.204	0.088	0.504	0.324
GraphRouter	0.316	0.602	0.290	0.222	0.170	0.084	0.416	0.300
Prompt LLM*	0.236	0.446	0.164	0.118	0.080	0.036	0.208	0.184
KNN Router*	0.202	0.398	0.166	0.096	0.060	0.030	0.176	0.161
Router-R1-Llama	**0.416**	**0.680**	**0.432**	**0.322**	**0.368**	0.128	0.520	**0.409**

Experiments

Methods	NQ† EM$^\uparrow$	NQ† Cost$^\downarrow$	PopQA EM$^\uparrow$	PopQA Cost$^\downarrow$	HpQA† EM$^\uparrow$	HpQA† Cost$^\downarrow$	2wiki EM$^\uparrow$	2wiki Cost$^\downarrow$
Qwen2.5-3B-Instruct								
Prompt LLM	0.300	20.0	0.340	17.6	0.268	20.1	0.262	20.2
Largest LLM	0.296	20.2	0.354	17.6	0.278	20.1	0.274	20.1
KNN Router	0.262	41.0	0.222	30.3	0.224	45.4	0.196	51.8
MLP Router	0.252	35.8	0.222	21.0	0.198	29.5	0.210	41.3
BERT Router	0.230	26.0	0.192	16.3	0.216	26.0	0.206	26.3
RouterDC	0.278	57.5	0.282	21.8	0.244	62.3	0.218	40.7
GraphRouter	0.276	29.6	0.280	19.2	0.234	24.7	0.180	28.6
Prompt LLM*	0.258	286.4	0.256	111.7	0.206	313.4	0.248	222.4
KNN Router*	0.236	102.2	0.232	49.8	0.154	133.0	0.234	99.4
Qwen2.5-7B-Instruct	0.138	24.2	0.130	18.6	0.152	24.9	0.166	28.7
LLaMA-3.1-70B-Instruct	0.280	153.3	0.342	76.3	0.260	124.6	0.270	119.8
LLaMA-3.1-8B-Instruct	0.242	29.5	0.208	14.3	0.198	24.1	0.130	21.3
Mistral-7B-Instruct	0.192	20.2	0.182	16.2	0.202	19.1	0.198	18.3
Mixtral-8x22B-Instruct	0.296	20.2	0.354	17.6	0.278	20.1	0.274	20.1
Gemma-2-27B-Instruct	0.282	42.7	0.290	22.0	0.244	37.8	0.190	44.6
Router-R1-Qwen ($\alpha = 0.0$)	**0.388**	150.6	0.384	98.3	**0.352**	138.6	**0.434**	150.8
Router-R1-Qwen ($\alpha = 0.6$)	**0.388**	150.0	**0.414**	75.9	0.332	124.3	0.422	113.8
Router-R1-Qwen ($\alpha = 0.7$)	0.318	32.3	0.280	17.2	0.254	27.2	0.236	31.4
Router-R1-Qwen ($\alpha = 0.8$)	0.320	28.9	0.270	14.9	0.238	28.2	0.202	31.4
Router-R1-Qwen ($\alpha = 0.9$)	0.244	**5.5**	0.234	**5.3**	0.186	**5.3**	0.252	**6.5**



THANK YOU

⭐ **Github: https://github.com/ulab-uiuc/Router-R1**

⭐ **Homepage: https://viktoraxelsen.github.io/**

Welcome Star & Discussion