FrugalRAG: Less is More in RL Finetuning for Multi-hop Question Answering

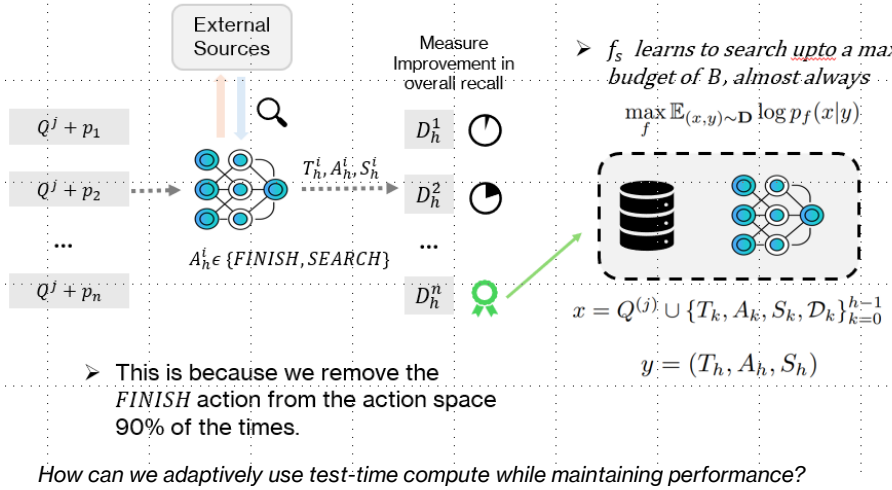
Abhinav Java, Srivathsan Koundinyan, Nagarajan Natarajan & Amit Sharma
Microsoft Research India

ICLR

Microsoft Research

Teaching models when to stop searching improves RL sample efficiency for training multi-hop RAG

Stage 1: *With sufficient test-time compute, LMs can perform multi-hop search and reasoning.*



- f_s *learns to search upto a maximum budget of* B*, almost always*

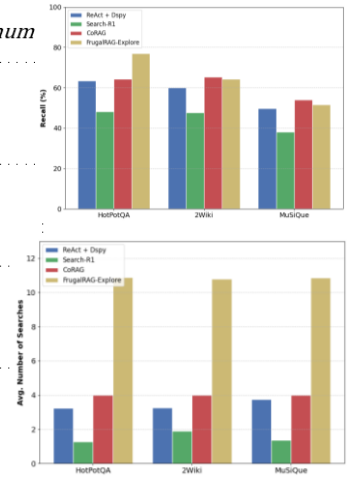
$$\max_f \mathbb{E}_{(x,y)\sim \mathbf{D}} \log p_f(x|y)$$

$$x = Q^{(j)} \cup \{T_k, A_k, S_k, \mathcal{D}_k\}_{k=0}^{h-1}$$

$$y = (T_h, A_h, S_h)$$

- This is because we remove the *FINISH* action from the action space 90% of the times.

How can we adaptively use test-time compute while maintaining performance?



Stage 2: *Learning nuanced decisions with a small action space improves RL sample efficiency.*

$$\mathbf{R} = \begin{cases} \max\left(-R_{\max}, \min\left(\log\left(\frac{1-\Delta}{\Delta}\right), R_{\max}\right)\right), & \text{if } \Delta > 0 \wedge c \geq \tau \quad \text{(late stop)} \\ R_{\max} + \alpha \cdot \left(\frac{h^*}{B}\right), & \text{if } \Delta = 0 \wedge c \geq \tau \quad \text{(perfect stop)} \\ \max\left(-R_{\max}, \min\left(\log\left(\frac{1-\Delta}{\Delta}\right), 0\right)\right), & \text{if } c < \tau \quad \text{(early stop)} \end{cases}$$

$$\Delta = \frac{|h_{term} - h^*|}{B}$$

Stop at Optimal Stopping Point h^*

Given $\boldsymbol{Q}$, let the maximum recall achieved by the base policy $\boldsymbol{f_s}$, be $\boldsymbol{\tau}$.

- Current recall $c \geq \tau$
- c does not improve if $h > h^*$

If $c < \tau$, then the model must *continue searching*

If $\Delta = 0$, then the model must get the maximum reward.

But two trajectories, both with $\Delta = 0$, must not get same reward unless they are of same length.

Comparison with State of the Art Multi-Hop RAG

Method	#Train	HotPotQA			2Wiki			MuSiQue		
		MBE	Recall	Searches	MBE	Recall	Searches	MBE	Recall	Searches
Zero-Shot	–	28.10	NA	NA	28.4	NA	NA	10.60	NA	NA
Zero-Shot CoT	–	29.10	NA	NA	29.8	NA	NA	10.7	NA	NA
Zero-Shot RAG	–	45.50	52.49	1	28.80	35.00	1	18.00	22.01	1
ReAct + DsPy	100	49.30	63.40	3.23	35.65	60.03	3.27	27.30	49.8	3.74
IRCOT	–	45.60	66.90	3.31	30.53	57.80	3.41	20.62	44.10	3.19
ITER-RETGEN	36k	46.10	55.40	3.00	32.10	45.30	3.00	19.10	34.20	3.00
Ret-Robust	1000	39.38	35.90	4.02	45.17	31.50	4.23	24.24	18.80	4.23
Self RAG 7B	150k	37.60	52.60	1.00	30.10	41.30	1.00	15.00	26.70	1.00
Self RAG 13B	150k	39.70	52.60	1.00	31.50	41.30	1.00	15.00	26.70	1.00
O2 Searcher	1440	42.70	50.10	1.77	45.70	55.20	2.42	26.40	37.00	1.95
SimpleDeepSearcher	871	50.40	64.80	2.75	49.30	60.50	3.64	34.30	50.40	2.86
R1-Searcher	8k	57.66	69.10	2.22	52.00	60.40	2.36	39.78	**57.70**	2.31
Search-R1	>100k	46.20	48.20	1.28	36.20	47.70	1.89	24.80	38.10	1.36
CoRAG	>100k	58.20	64.30	4.00	**59.00**	**65.40**	4.00	**40.50**	54.00	4.00
FRUGALRAG-7B *+Qwen2.5-7B-Inst*	1000	58.5	**70.40**	2.89	50.40	58.80	3.03	36.40	53.30	3.30
FRUGALRAG-7B *+ Qwen2.5-32B-Inst*	1000	**61.4**	**70.40**	2.89	50.90	58.80	3.03	39.90	53.30	3.30
FRUGALRAG-7B *+ CoRAG*	1000	58.00	**70.40**	2.89	51.20	58.80	3.03	34.60	53.30	3.30

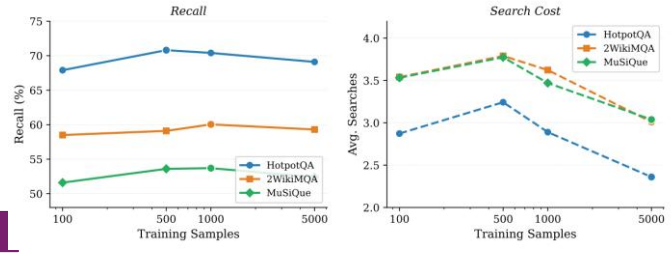
Generalization to Deep Research on unseen corpus

Method	Model Size	Accuracy (%)	Recall (%)	Avg. Searches
Sonnet 4	-	37.35	47.33	9.03
Opus 4	-	36.75	50.84	10.24
kimi-k2-0711-preview	-	35.42	38.38	11.22
Gemini-2.5-Flash	-	34.58	40.19	9.77
Gemini-2.5-Pro	-	29.52	35.31	6.04
oss-120b-low	120B	25.54	22.50	2.21
oss-20b-high	20B	35.06	49.29	23.87
oss-20b-medium	20B	30.48	41.31	13.64
oss-20b-low	20B	14.10	17.37	1.87
DeepSeek-R1-0528	600B	16.39	16.32	2.72
Qwen3-32B	32B	10.72	7.80	0.94
Search-R1-32B	32B	11.08	10.17	1.69
FRUGALRAG-7B + *Qwen2.5-7B-Inst* (HotPotQA)	7B	20.46	23.57	7.95
FRUGALRAG-7B + *Qwen2.5-7B-Inst* (2Wiki)	7B	21.53	22.93	10.96
FRUGALRAG-7B + *Qwen2.5-7B-Inst* (MuSiQue)	7B	21.14	23.73	8.39

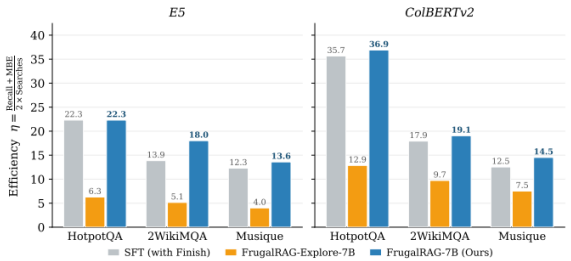
FrugalRAG *learns to use more test-time compute on harder questions and becomes for efficient with data*

2WikiMultiHopQA			MuSiQue		
Num GT Evidence	Number of Questions	Searches	Actual Hops	Number of Questions	Searches
2	9,595	2.665 ± 1.430	2	1,252	3.054 ± 1.433
3	88	2.875 ± 1.483	3	760	3.924 ± 1.464
4	2,806	3.909 ± 1.502	4	405	4.205 ± 1.368

Table 4: FRUGALRAG issues more queries for harder questions. We see a clear increasing trend in num. of search queries with amount of ground truth evidence (2Wiki) and ground truth num. of hops.



Reward Shaping



*OOD Generalization**

	HotPotQA			2Wiki			MuSiQue		
Method Variant	Recall	MBE	Searches	Recall	MBE	Searches	Recall	MBE	Searches
FrugalRAG + E5 (HotPotQA)	70.40	69.50	2.89	60.40	49.66	3.623	53.70	34.10	3.47
FrugalRAG + E5 (2Wiki)	71.00	54.96	2.85	58.80	53.00	3.03	52.22	29.41	3.38
FrugalRAG + E5 (MuSiQue)	69.10	53.95	2.72	59.70	52.20	3.33	53.30	41.50	3.30
FrugalRAG + ColBERTv2 (HotPotQA)	82.80	68.47	2.05	64.20	48.47	3.07	53.80	36.27	2.75
FrugalRAG + ColBERTv2 (2Wiki)	81.90	68.34	2.56	63.50	48.93	2.95	53.80	34.30	3.10
FrugalRAG + ColBERTv2 (MuSiQue)	83.10	61.11	2.53	60.80	46.41	3.13	51.70	29.80	3.02