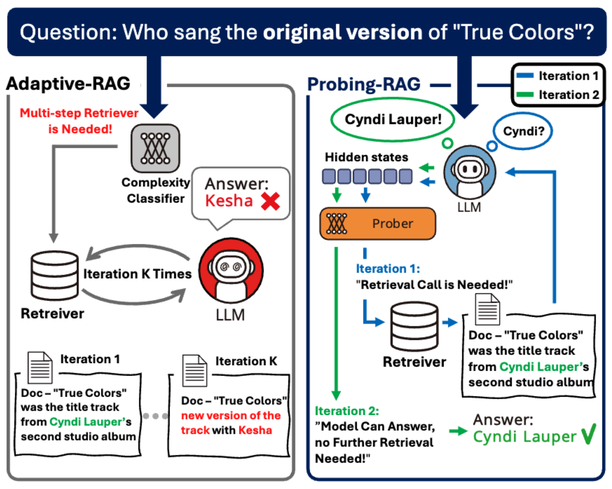
Probing-RAG: Self-Probing to Guide Language Models in Selective Document Retrieval

Ingeol Baek, Hwan Chang, Byeongjeong Kim, Jimin Lee, Hwanhee Lee

Language Intelligence lab, Chung-Ang University

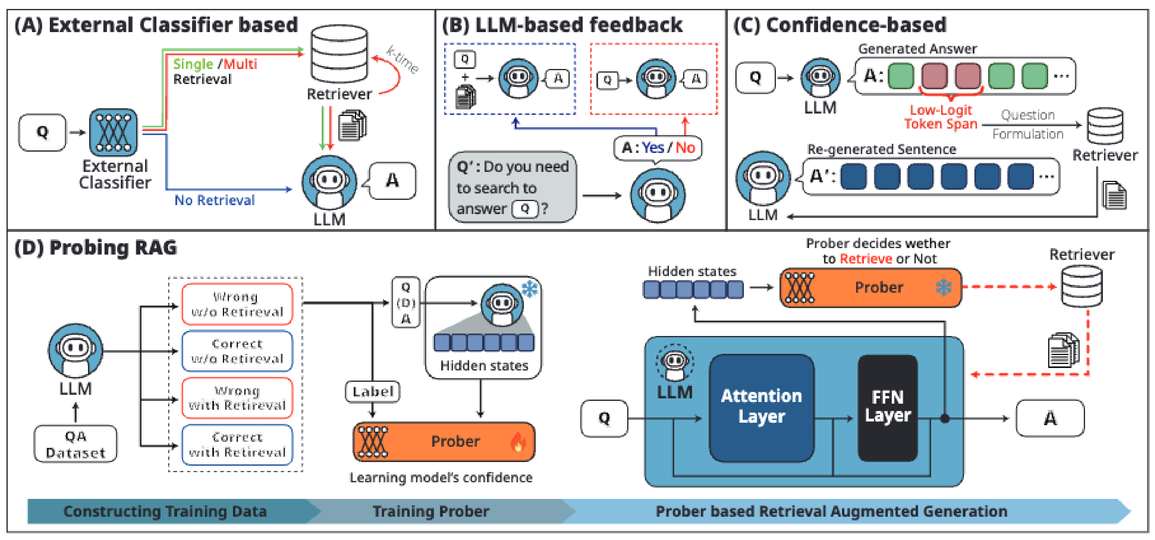
NAACL 2025

Motivation

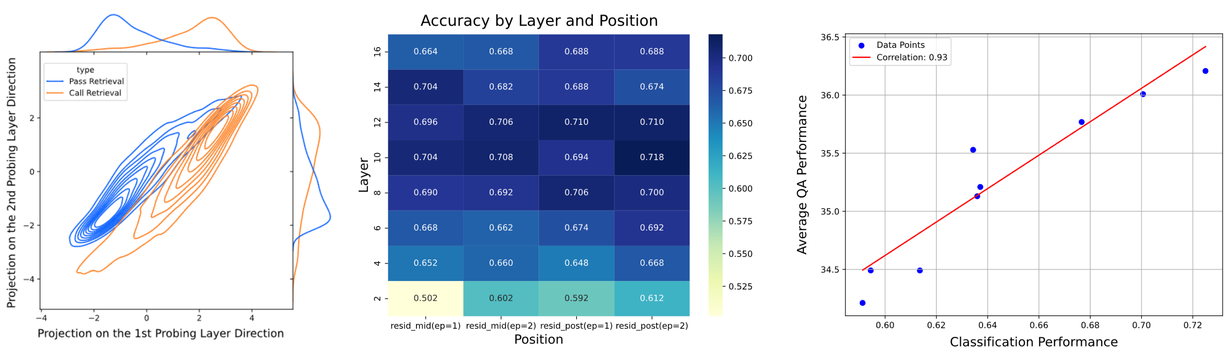


- Recently, research on Retrieval-Augmented Generation (RAG) has been actively exploring Adaptive-RAG, which **dynamically determines whether additional information retrieval is needed**.
- Previous Adaptive-RAG approaches employed an external classifier to decide how many retrievals were necessary. **However, this method failed to consider the model's internal knowledge**.
- Furthermore, recent LLMs possess extensive knowledge and **may be capable of generating responses without the need for search**. However, we intentionally disregard this ability and employ RAG, which **can lead to conflicts between the model's internal knowledge and the information retrieved from documents**.

Method



- we use a module called **Prober to determine whether additional retrieval is needed**.
- The Prober module is a binary classification module that takes the hidden state as input, after being integrated with the model's internal residual network.



- Using the two Probers that showed the best performance, we extracted the kernel density estimation, and **notable differences were observed**.
- Additionally, we confirmed that **the accuracy of the Prober exceeded 0.7 for each layer**.
- Finally, as the performance of the Prober improved, we observed a **corresponding increase in QA performance, with a high correlation of 0.93**.

Training Dataset Construction

	No Retrieval	With Retrieval
Correct	**Query:** What race track in the midwest hosts a 500 mile race every May? **Rationale:** The Indianapolis 500 is a 500-mile (800 km) automobile race... **Pred:** Indianapolis Motor Speedway **Label:** Indianapolis Motor Speedway **ACC (Prober Training Label): 1**	**Query:** Were Scott Derrickson and Ed Wood of the same nationality? **Passages:** {Retrieved passage 1, ..., Retrieved passage 5} **Rationale:** Scott Derrickson is an American film director, screenwriter... **Pred:** Yes **Label:** Yes **ACC (Prober Training Label): 1**
Incorrect	**Query:** When was Poison's album "Shut Up, Make Love" released? **Rationale:** The album "Shut Up, Make Love" was released in ... **Pred:** 1990 **Label:** 2000 **ACC (Prober Training Label): 0**	**Query:** Are Random House Tower and 888 7th Avenue both used for real estate? **Passages:** {Retrieved passage 1, ..., Retrieved passage 5} **Rationale:** Random House Tower is a 40-story skyscraper in... **Pred:** Yes **Label:** No **ACC (Prober Training Label): 0**

- To train the Prober module, we used datasets like HotpotQA, NaturalQA, and TriviaQA, **creating four types of datasets**.

Experiment

Methods	In-Domain HotpotQA EM	In-Domain HotpotQA ACC	In-Domain NQ EM	In-Domain NQ ACC	In-Domain TriviaQA EM	In-Domain TriviaQA ACC	Out-of-Domain MuSiQue EM	Out-of-Domain MuSiQue ACC	Out-of-Domain 2Wiki EM	Out-of-Domain 2Wiki ACC	Average EM	Average ACC
Gemma-2b												
No Retrieval	16.8	28.0	15.0	24.6	37.4	45.4	3.2	4.8	22.6	43.0	19.0	29.2
Single-step Approach	14.6	28.2	11.4	26.0	19.6	38.8	1.8	5.8	22.8	38.4	14.0	27.4
LLM-based	18.6	25.8	17.6	20.4	36.8	41.8	3.8	4.6	24.2	25.8	20.2	23.7
FLARE	13.2	21.0	9.0	21.8	13.8	31.0	1.2	5.0	21.6	27.8	11.8	21.3
DRAGIN	19.8	22.6	18.8	22.2	**42.8**	47.0	4.2	4.8	**26.4**	28.8	22.4	25.1
Adaptive-RAG	13.2	23.6	11.4	26.2	22.8	40.8	1.2	3.0	21.6	40.6	14.0	26.8
Probing-RAG(Ours)	**21.8**	**39.4**	**21.6**	**35.0**	41.8	**52.2**	**4.8**	**8.8**	24.2	**43.6**	**22.8**	**35.8**
Mistral-7b												
No Retrieval	17.0	20.6	13.2	19.8	38.0	45.2	3.4	6.2	16.4	30.0	17.6	24.4
Single-step Approach	18.6	34.2	16.8	35.0	34.6	51.0	5.4	9.0	21.6	32.6	19.4	32.4
LLM-based	20.4	32.0	15.8	35.6	41.2	49.8	5.6	9.4	16.8	32.4	20.0	31.8
FLARE	20.4	32.0	15.4	34.4	35.0	45.6	4.4	6.6	18.6	31.0	18.8	29.9
DRAGIN	21.2	28.0	16.8	37.2	39.8	42.2	5.2	7.2	**23.2**	25.8	21.2	28.1
Adaptive-RAG	19.0	26.0	17.2	37.4	40.8	50.2	4.0	5.8	22.6	31.6	20.7	30.2
Probing-RAG(Ours)	**22.4**	**38.6**	**20.8**	**39.4**	**43.2**	**52.2**	**5.8**	**9.8**	23.0	**33.4**	**23.0**	**34.7**

- In the experimental results, we were able to achieve the best performance with the Prober trained on both **in-domain and out-of-domain datasets**.

	Total Retrieval Call	No Retrieval Step Ratio	Single-step Retrieval Step Ratio	Multi-step Retrieval Step Ratio
LLM-based	2345	6.2%	93.8%	0.00%
FLARE	5317	12.41%	29.35%	58.24%
DRAGIN	13570	0.00%	1.20%	98.80%
Adaptive-RAG	3068	7.79%	61.96%	30.25%
Probing-RAG	1988	57.46%	20.19%	22.35%

- We measured the consistency of each approach by evaluating how consistently they matched the queries answered correctly during DirectQA.

Consistency	HotpotQA	NQ	TriviaQA	MuSiQue	2Wiki
No Retrieval	100%	100%	100%	100%	100%
Single-step-RAG	73.8%	76.0%	82.0%	70.4%	85.0%
FLARE	74.4%	76.4%	83.8%	68.8%	72.8%
DRAGIN	75.2%	75.0%	81.8%	77.8%	69.8%
Adaptive-RAG	83.0%	76.6%	86.0%	74.6%	93.2%
Probing-RAG	90.6%	92.6%	91.0%	96.4%	96.6%

- While all methods showed a relatively low consistency of about 70%, Probing-RAG demonstrated a consistency of around 90%.