Beyond RAG vs. Long-Context: Learning Distraction-Aware Retrieval for Efficient Knowledge Grounding

Seong-Woong Shim*, Myunsoo Kim*, Jae Hyeon Cho, Byung-Jun Lee

ICLR
International Conference On Learning Representations

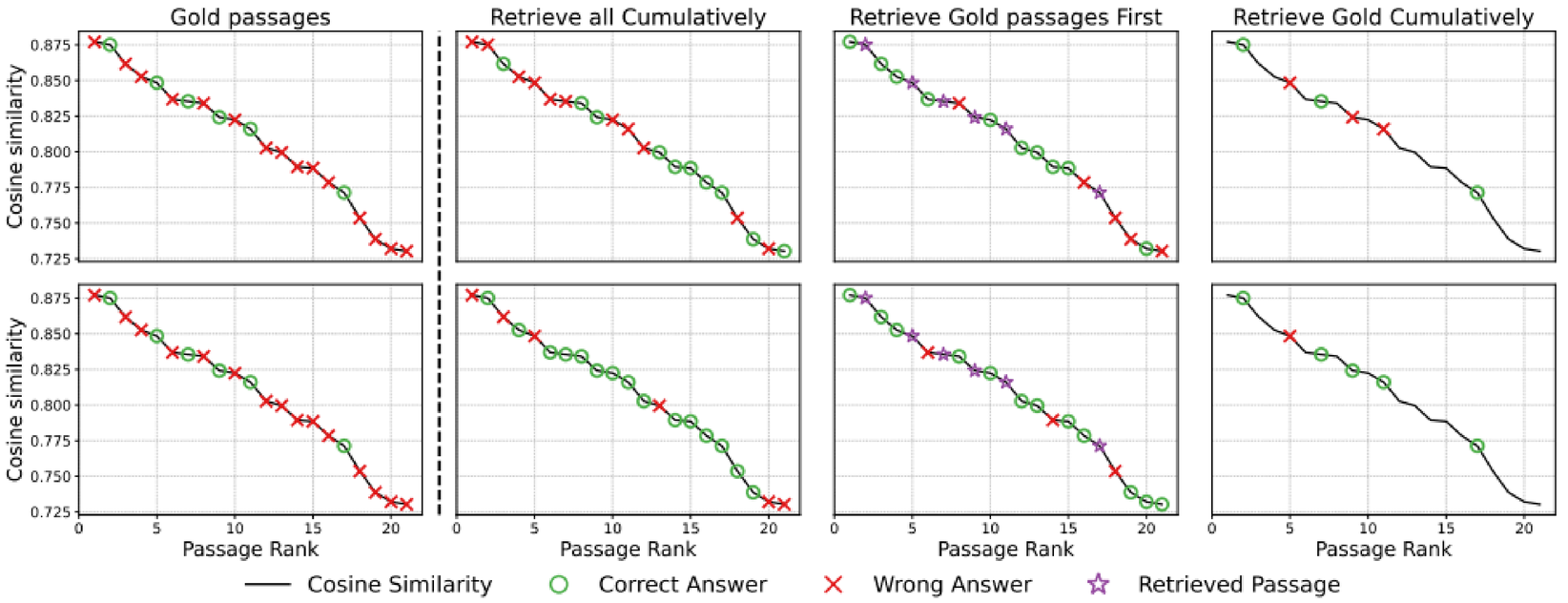
Motivation

- Although prior works have highlighted the detrimental impact of distracting passages on retrieval performance, relatively little attention has been paid to retrieval strategies that explicitly mitigate such influence.

- We define **distracting passages** as passages that misguide the LLM in generating the correct answer, irrespective of whether they lead to correct.

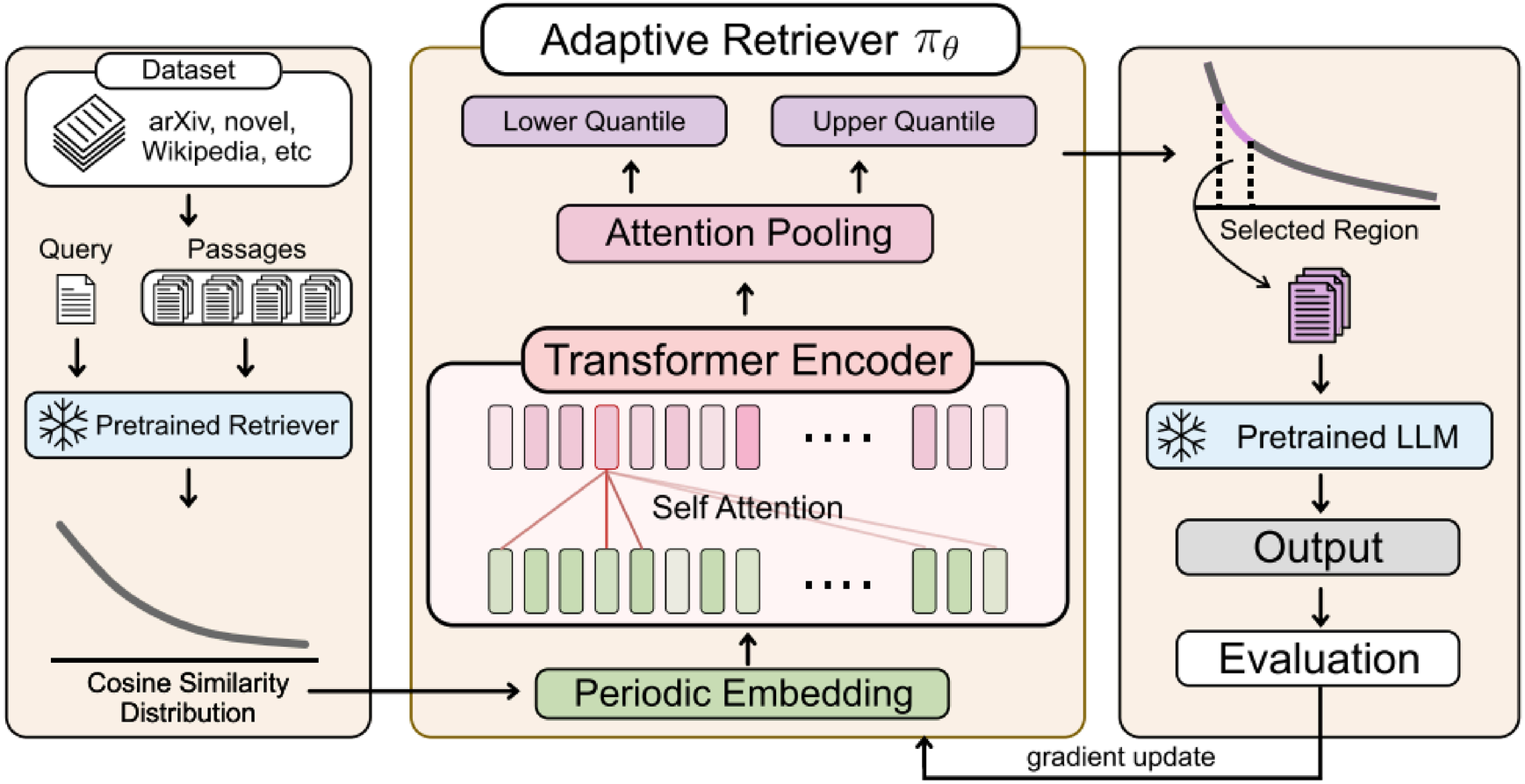
[Open-source]
Llama-3.1-8B

[Close-source]
GPT-4o



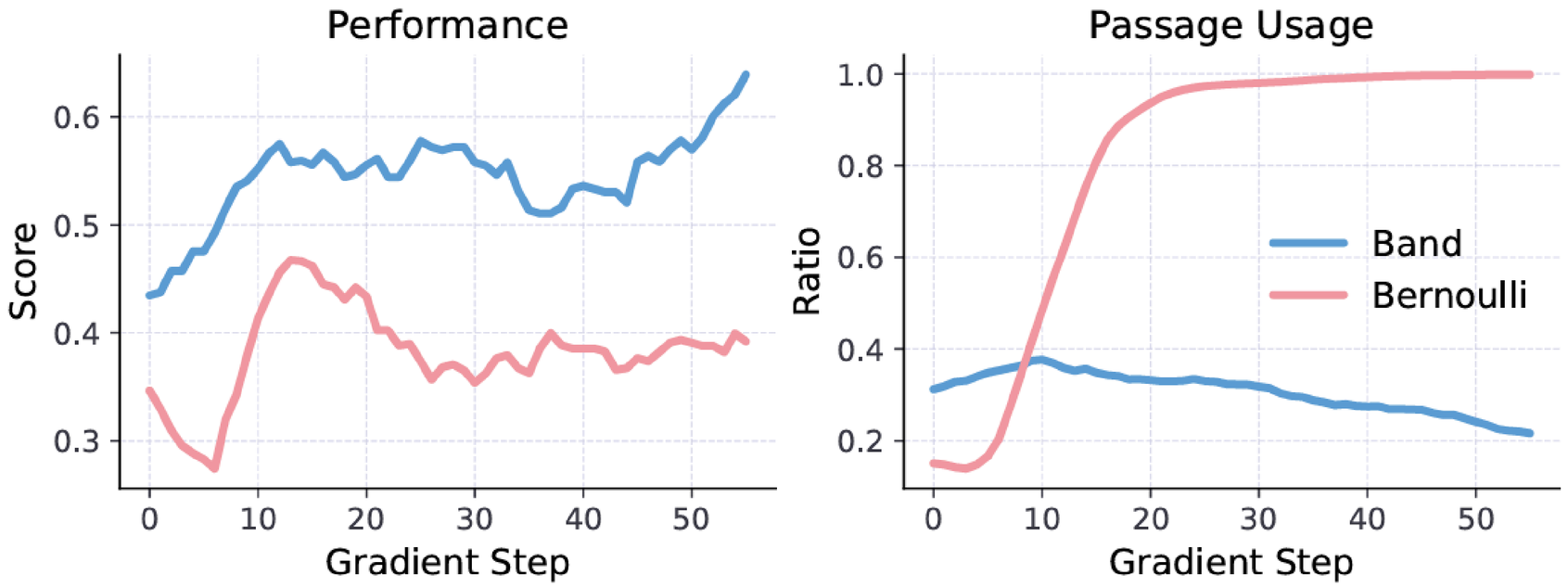
Method

- Overview of LDAR, a learning-based retrieval strategy that adapts to each LLM by balancing information coverage and distraction.



Method

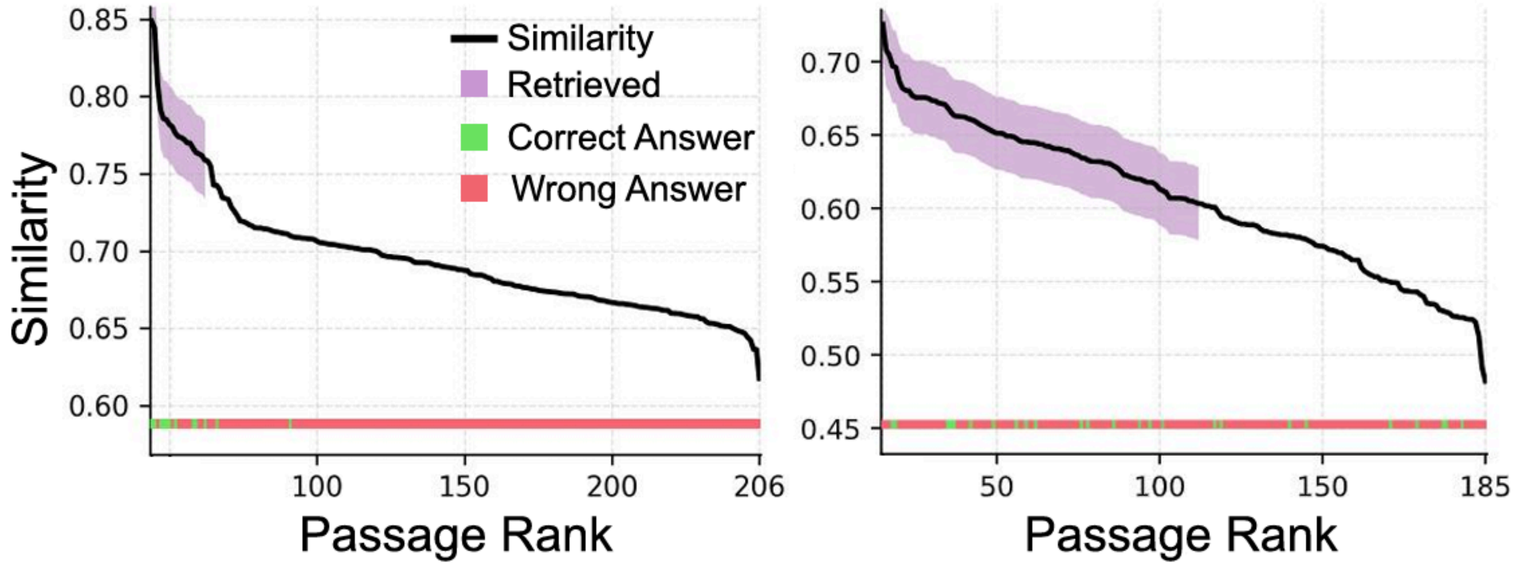
- A naive Bernoulli-based apporach fails to identify a balanced trade-off between the RAG and long-context approach.
- This limitation arises from its need to explore the entire combinatorial subset selection space, which impedes generalization and ultimately causes convergence to a local optimum (corresponding to the long-context approach in this case).



Experiment

- LDAR adapts its retrieval strategy based on the similarity distribution between the query and passages.

- If passages with high semantic similarity exist, it tends to focus narrowly on that region. In contrast, when no passages exhibit strong semantic similarity, it expands the retrieval range to ensure broader information coverage, even at the cost of incorporating more potential distracting passages.



Experiment

- LDAR generally achieves significantly higher performance compared to all other baselines, while using only about half the token usage of the long-context approach.

- The average token usage ratio of LDAR relative to the long-context approach is: 0.47 (32K opensource), 0.63 (32K closed-source), 0.25 (128K open-source), 0.52 (128K closed-source).

Method	Location		Reasoning		Comparison		Hallucination		Overall		Location		Reasoning		Comparison		Hallucination		Overall	
	Context Length 32k										*Context Length 128k*									
Top-1	52.7± 0.5 (0.019)	56.6± 0.9 (0.019)	23.3± 3.2 (0.019)	37.1± 2.2 (0.019)	20.6± 2.2 (0.018)	33.0± 3.6 (0.018)	**86.0**± 3.1 (0.019)	**89.4**± 2.8 (0.019)	45.65± 2.3 (0.018)	54.02± 2.4 (0.018)	31.1± 1.1 (0.005)	31.5± 0.9 (0.005)	26.1± 2.8 (0.004)	33.7± 2.4 (0.004)	4.30± 1.5 (0.005)	12.1± 4.4 (0.005)	**84.9**± 4.5 (0.005)	**83.7**± 4.0 (0.005)	36.60± 2.5 (0.004)	40.25± 2.9 (0.004)
Top-5	66.7± 1.2 (0.095)	78.0± 0.5 (0.095)	38.2± 4.8 (0.097)	61.6± 4.8 (0.097)	47.6± 2.7 (0.091)	62.8± 3.8 (0.091)	82.5± 3.4 (0.095)	84.3± 3.3 (0.095)	58.75± 3.0 (0.094)	71.67± 3.1 (0.094)	60.1± 1.4 (0.026)	61.7± 1.8 (0.026)	47.4± 3.6 (0.024)	62.6± 3.3 (0.024)	24.3± 3.0 (0.025)	34.1± 2.6 (0.025)	74.1± 6.6 (0.026)	71.6± 5.4 (0.026)	51.47± 3.7 (0.025)	57.50± 3.3 (0.025)
Top-10	75.3± 1.6 (0.190)	83.4± 1.1 (0.190)	41.6± 4.0 (0.194)	59.5± 3.0 (0.194)	51.5± 5.9 (0.182)	65.2± 3.6 (0.182)	79.1± 6.1 (0.190)	80.8± 4.3 (0.190)	61.87± 4.4 (0.189)	72.22± 3.0 (0.189)	66.9± 0.6 (0.053)	70.8± 2.3 (0.053)	53.7± 3.7 (0.049)	67.5± 2.3 (0.049)	29.2± 5.1 (0.050)	42.6± 3.2 (0.050)	70.3± 0.9 (0.052)	64.3± 7.1 (0.052)	55.02± 2.6 (0.051)	61.30± 3.7 (0.051)
Top-25	78.1± 3.0 (0.474)	87.4± 1.2 (0.474)	39.9± 4.0 (0.486)	61.6± 0.9 (0.486)	50.2± 6.9 (0.457)	70.9± 3.7 (0.457)	74.4± 8.6 (0.476)	77.0± 4.8 (0.476)	60.65± 5.6 (0.473)	74.22± 2.7 (0.473)	71.3± 2.7 (0.133)	80.0± 1.2 (0.133)	54.5± 4.6 (0.124)	75.3± 1.9 (0.124)	24.8± 2.2 (0.126)	48.1± 2.5 (0.126)	65.1± 12.0 (0.131)	55.6± 7.8 (0.131)	53.92± 5.4 (0.128)	64.75± 3.4 (0.128)
Top-50	78.1± 2.8 (0.866)	87.0± 2.5 (0.866)	37.4± 4.0 (0.897)	63.8± 2.6 (0.897)	49.6± 4.2 (0.853)	69.3± 5.4 (0.853)	72.1± 10.6 (0.862)	74.7± 5.0 (0.862)	59.30± 5.4 (0.869)	73.70± 3.9 (0.869)	67.9± 4.1 (0.267)	80.3± 0.5 (0.267)	52.4± 3.9 (0.249)	74.3± 2.0 (0.249)	33.6± 3.6 (0.252)	52.4± 4.7 (0.252)	61.2± 11.8 (0.262)	52.7± 7.4 (0.262)	53.77± 5.9 (0.257)	64.92± 3.7 (0.257)
LC	80.3± 1.9 (1.000)	87.4± 0.7 (1.000)	36.9± 4.0 (1.000)	62.2± 3.2 (1.000)	47.7± 5.3 (1.000)	73.5± 4.0 (1.000)	69.6± 8.4 (1.000)	73.0± 7.3 (1.000)	58.62± 4.9 (1.000)	74.00± 3.8 (1.000)	56.2± 10.1 (1.000)	88.0± 2.3 (1.000)	45.1± 6.4 (1.000)	76.9± 2.2 (1.000)	23.8± 5.9 (1.000)	56.6± 4.0 (1.000)	49.5± 11.3 (1.000)	58.4± 9.8 (1.000)	43.65± 8.4 (1.000)	69.97± 4.6 (1.000)
RAG	76.3± 1.7 (0.095)	82.5± 0.5 (0.095)	43.7± 3.2 (0.097)	59.0± 1.3 (0.097)	47.0± 6.5 (0.091)	61.2± 5.7 (0.091)	81.5± 3.7 (0.095)	79.8± 4.7 (0.095)	62.12± 3.8 (0.094)	70.62± 3.1 (0.094)	71.6± 1.5 (0.026)	75.2± 2.1 (0.026)	49.8± 3.6 (0.024)	65.8± 2.3 (0.024)	26.4± 4.4 (0.025)	42.6± 2.3 (0.025)	70.9± 6.5 (0.026)	74.1± 6.5 (0.026)	54.67± 4.0 (0.025)	64.42± 3.3 (0.025)
Self-Route	80.6± 1.7 (0.255)	89.6± 0.8 (0.295)	40.3± 5.0 (0.258)	62.7± 4.1 (0.232)	47.0± 4.6 (0.244)	67.6± 2.9 (0.312)	69.9± 8.2 (0.949)	76.0± 5.2 (0.967)	59.45± 4.9 (0.426)	73.97± 3.3 (0.451)	65.8± 2.6 (0.187)	80.0± 3.2 (0.282)	51.3± 4.6 (0.181)	59.7± 2.0 (0.249)	25.3± 0.2 (0.381)	52.4± 5.0 (0.625)	56.4± 11.3 (0.888)	57.5± 9.7 (0.943)	49.70± 4.7 (0.409)	62.40± 5.0 (0.524)
Adaptive-k	61.0± 2.2 (0.395)	71.4± 2.2 (0.395)	25.4± 0.4 (0.362)	51.5± 3.5 (0.362)	43.1± 3.0 (0.385)	59.6± 3.8 (0.385)	77.3± 7.1 (0.479)	82.2± 4.4 (0.479)	51.70± 3.2 (0.405)	66.17± 3.5 (0.405)	47.7± 4.7 (0.398)	64.4± 2.3 (0.398)	32.1± 6.0 (0.405)	54.1± 1.4 (0.405)	18.5± 3.0 (0.675)	44.4± 1.8 (0.675)	66.0± 8.8 (0.502)	68.3± 6.8 (0.502)	41.07± 5.6 (0.495)	57.80± 3.1 (0.495)
BGM	78.8± 0.9 (0.048)	82.6± 0.8 (0.057)	46.5± 4.0 (0.067)	59.0± 1.6 (0.074)	50.1± 1.6 (0.066)	61.2± 6.9 (0.064)	75.4± 4.7 (0.049)	75.7± 2.8 (0.045)	62.70± 2.8 (0.057)	69.63± 3.0 (0.060)	68.9± 1.6 (0.017)	72.2± 3.0 (0.019)	56.5± 3.1 (0.013)	66.8± 0.8 (0.017)	30.3± 4.8 (0.015)	34.2± 2.2 (0.020)	73.1± 6.1 (0.015)	73.4± 6.0 (0.016)	57.20± 3.9 (0.015)	61.65± 3.0 (0.018)
RankZephyr	72.4± 1.2 (0.095)	82.1± 1.2 (0.095)	36.9± 4.7 (0.097)	59.5± 2.7 (0.097)	44.4± 4.3 (0.091)	62.1± 4.8 (0.091)	80.5± 4.4 (0.095)	85.2± 2.4 (0.095)	58.55± 3.7 (0.094)	72.2± 2.8 (0.094)	56.1± 2.1 (0.026)	62.7± 0.9 (0.026)	54.5± 3.3 (0.024)	71.4± 1.9 (0.024)	20.9± 4.5 (0.025)	28.0± 3.6 (0.025)	73.3± 5.9 (0.026)	75.8± 6.5 (0.026)	51.20± 4.0 (0.025)	59.48± 3.2 (0.025)
LDAR	**87.7**± 1.4 (0.478)	**91.9**± 1.2 (0.628)	**52.7**± 4.7 (0.400)	**70.1**± 0.9 (0.636)	**63.1**± 2.6 (0.518)	**78.9**± 4.0 (0.619)	76.5± 6.9 (0.474)	76.8± 5.8 (0.633)	**70.00**± 3.9 (0.467)	**79.42**± 3.0 (0.629)	**77.3**± 2.2 (0.209)	**90.5**± 0.8 (0.502)	**61.7**± 2.3 (0.272)	**82.7**± 0.8 (0.444)	**42.9**± 2.3 (0.312)	**65.8**± 1.4 (0.606)	64.3± 10.5 (0.209)	65.9± 6.3 (0.519)	**61.55**± 4.3 (0.250)	**76.22**± 2.3 (0.517)

Experiment

- In zero-shot setting, LDAR still achieves better average performance compared to RAG or the long-context approach, while also attaining a lower token usage ratio relative to the long-context approach

Method	HotpotQA			NQ		
Llama-3.1-8B-Instruct	52.0 (1.000)	50.0 (0.019)	**59.0** (0.499)	43.0 (1.000)	40.0 (0.021)	**49.0** (0.213)
Llama-3.2-3B-Instruct	**54.0** (1.000)	52.0 (0.019)	**54.0** (0.207)	42.0 (1.000)	37.0 (0.021)	**43.0** (0.146)
Qwen-2.5-7B-Instruct	30.0 (1.000)	63.0 (0.019)	**64.0** (0.305)	25.0 (1.000)	42.0 (0.021)	**54.0** (0.126)
Qwen-3-4B-Instruct	56.0 (1.000)	51.0 (0.019)	**62.0** (0.536)	**54.0** (1.000)	41.0 (0.021)	53.0 (0.486)
Mistral-Nemo-12B	29.0 (1.000)	**63.0** (0.019)	61.0 (0.061)	23.0 (1.000)	45.0 (0.021)	**47.0** (0.099)
Open-source Average	44.2 (1.000)	55.8 (0.019)	**60.0** (0.321)	37.4 (1.000)	41.0 (0.021)	**49.2** (0.214)
GPT-4o	81.0 (1.000)	65.0 (0.019)	**84.0** (0.579)	**61.0** (1.000)	54.0 (0.021)	60.0 (0.738)
GPT-4o-mini	64.0 (1.000)	65.0 (0.019)	**76.0** (0.629)	**59.0** (1.000)	52.0 (0.021)	**59.0** (0.374)
Gemini-2.5-pro	**85.0** (1.000)	55.0 (0.019)	84.0 (0.638)	62.0 (1.000)	37.0 (0.021)	**65.0** (0.518)
Gemini-2.5-flash	82.0 (1.000)	57.0 (0.019)	**83.0** (0.953)	54.0 (1.000)	37.0 (0.021)	**60.0** (0.564)
Closed-source Average	78.0 (1.000)	60.5 (0.019)	**81.8** (0.699)	59.0 (1.000)	45.0 (0.021)	**61.0** (0.457)

Thank you