Task	NQ	TriviaQA	PopQA	HotpotQA	2WikimQA	FEVER	Doc2Dial	TopiOCQA	Inscit
Metric	EM	EM / Acc.	EM / Acc.	EM / F1	EM / F1	Acc.	F1	F1	F1
RankRAG 8B v.s. ChatQA-1.5 8B	<1e-4	<1e-4 / <1e-4	<1e-4 / <1e-4	2e-4 / <1e-4	<1e-4 / <1e-4	<1e-4	1e-4	0.062	<1e-4
RankRAG 70B v.s. ChatQA-1.5 70B	<1e-4	4e-3 / 2e-3	<1e-4 / <1e-4	0.044 / 0.038	<1e-4 / <1e-4	<1e-4	0.151	–	<1e-4

Table 1: The p-value of fishers' randomization test on RAG tasks. Dark blue and blue are highly statistical significant and statistical significant results.

Task	NQ			TriviaQA			PopQA			HotpotQA			Inscit		
Recall	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20
RankRAG 8B v.s. RankLlama 8B	<1e-4	<1e-4	0.083	<1e-4	<1e-4	<1e-4	<1e-4	<1e-4	<1e-4	0.042	—	0.062	<1e-4	<1e-4	<1e-4

Table 2: The p-value of fishers' randomization test on ranking tasks. Dark blue and blue are highly statistically significant and statistically significant results.

Task (*Zero-shot*)	NQ	TriviaQA	PopQA	HotpotQA	2WikimQA	FEVER	Doc2Dial	TopiOCQA	Inscit
Metric	EM	EM / Acc.	EM / Acc.	EM / F1	EM / F1	Acc.	F1	F1	F1
Llama3-ChatQA-1.5 8B	42.4	81.0 / 87.6	52.6 / 59.8	33.4 / 44.6	26.8 / 31.9	90.9	39.3	49.9	30.1
Llama3-RankRAG 8B (top 30)	49.4†	82.2† / 89.0†	56.6† / 63.1†	34.2† / 45.6†	28.9† / 35.0†	91.7†	40.0‡	50.1	32.7†

Table 3: Result of RankRAG 8B using top 30 passages for reranking. † and ‡ are significant results with p<0.01/0.05.

Task (*Zero-shot*)	NQ	TriviaQA	PopQA	HotpotQA	2WikimQA	FEVER	Doc2Dial	TopiOCQA	Inscit
Model (Generation + Ranking) / Metric	EM	EM / Acc.	EM / Acc.	EM / F1	EM / F1	Acc.	F1	F1	F1
Llama3-Instruct 8B + RankLlama 8B	34.5	74.2 / 84.2	39.0 / 58.6	30.4 / 39.6	12.5 / 28.7	90.0	34.4	45.8	33.0
ChatQA-1.5 8B + RankLlama 8B	46.9	81.9 / 88.0	55.2 / 62.4	34.6 / 45.9	29.8 / 34.6	91.4	39.6	**50.6**	31.0
RankRAG 8B + RankLlama 8B	49.0	82.0 / 88.4	56.1 / 62.8	35.0 / 46.2	30.6 / 35.5	91.7	40.0	49.8	32.4
RankRAG 8B + RankRAG 8B	**50.6**†	**82.9**† **/ 89.5**†	**57.6**† **/ 64.1**†	**35.3 / 46.7**‡	**31.4**† **/ 36.9**†	**92.0**	**40.4**‡	50.4	**33.3**†

Table 4: Results from RankRAG and baseline models using RankLlama 8B as the reranker across 9 datasets. Baseline models rerank the top 100 passages and **have the same time efficiency** with RankRAG. † and ‡ are significant results with p<0.01/0.05.

Task (*Zero-shot*)	NQ	TriviaQA	PopQA	HotpotQA	2WikimQA	FEVER	Doc2Dial	TopiOCQA	Inscit
Metric	EM	EM / Acc.	EM / Acc.	EM / F1	EM / F1	Acc.	F1	F1	F1
GPT-3.5-turbo-1106	38.6	82.9 / 91.7	28.4 / 32.2	29.9 / 42.0	23.9 / 30.4	82.7	20.1	28.5	27.2
GPT-4-0613	40.3	84.8 / **94.5**	31.3 / 34.8	34.5 / 46.9	29.8 / 36.6	87.7	27.6	30.1	27.0
GPT-4-turbo-2024-0409	41.5	80.0 / 94.3	25.0 / 33.5	26.6 / 43.8	24.1 / 35.5	87.0	27.6	26.4	24.4
GPT-3.5-turbo-1106 RAG	46.7	79.7 / 88.0	49.9 / 57.0	31.2 / 41.2	27.2 / 32.2	90.8	34.8	44.3	35.3
GPT-4-0613 RAG†	40.4	75.0 / 88.5	44.3 / 61.4	27.6 / 38.1	14.4 / 17.6	92.6	34.2	45.1	**36.4**
GPT-4-turbo-2024-0409 RAG†	40.3	70.2 / 91.1	39.5 / 58.4	8.1 / 17.9	22.8 / 39.2	92.2	35.4	48.3	33.8
RankRAG 8B	50.6	82.9 / 89.5	57.6 / 64.1	35.3 / 46.7	31.4 / 36.9	92.0	40.4	50.4	33.3
RankRAG 70B	**54.2**	**86.5** / 92.3	**59.9 / 65.4**	**42.7 / 55.4**	**38.2 / 43.9**	**93.8**	**41.5**	**52.8**	35.2

Table 5: Results of RankRAG and OpenAI GPT-series model on 9 datasets. Note that†: GPT-4 and GPT-4-turbo may refuse to answer the question when retrieved passages do not contain relevant information, thus the EM / accuracy drops after including RAG on TriviaQA, HotpotQA, and 2WikimQA. **This set of experiments costs more than $1000**.

Task	# Data	NQ			TriviaQA			PopQA			HotpotQA			Inscit		
Recall		R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20	R@5	R@10	R@20
RankVicuna (**top 100**)	∼100k†	75.3	80.6	84.7	90.9	93.8	96.2	80.8	83.4	85.5	55.2	60.6	63.8	53.4	59.8	69.2
GPT-3.5 (**top 100**)	Unk.	77.8	82.5	85.7	91.1	94.4	96.7	77.4	82.0	85.5	52.1	56.6	62.4	50.2	59.1	68.6
GPT-4‡ (**top 30**)	Unk.	79.3	83.2	85.1	92.8	95.5	96.8	79.3	83.6	86.2	53.2	57.0	61.0	52.3	61.7	70.0
RankRAG 8B (**top 100**)	∼50k	80.3	**84.0**	**86.3**	93.2	95.4	**97.3**	81.6	**84.9**	**87.0**	**57.6**	61.8	**65.2**	60.9	65.7	73.5
RankRAG 70B (**top 30**)	∼50k	**80.6**	**84.0**	85.4	**93.6**	**95.9**	97.1	**81.8**	84.6	86.5	56.3	59.7	62.2	**61.3**	**66.4**	**74.6**

Table 6: Ranking performance with different ranking models. RankRAG achieves better performance despite using fewer ranking data. ‡ We only rerank top-30 passages for GPT-4 due to budget constraints. **This set of experiments costs more than $1500**.

	RankRAG w/ $k = 5$	RankRAG w/ adaptive k
NQ (EM)	50.6	**50.8**
HotpotQA (EM/F1)	**35.3 / 46.7**	33.5 / 45.3
FEVER (Acc.)	**92.0**	91.6
Inscit (F1)	**33.3**	32.7

Table 7: The performance comparison on RankRAG 8B.

	TriviaQA	PopQA
Self-RAG 13B w/ Contriever	69.3	55.8
RankRAG 8B w/ Contriever	77.3	57.3
Self-RAG 13B w/ Dragon	83.8	57.0
RankRAG 8B w/ Dragon	89.5	64.1

Table 8: Accuracy of Self-RAG and RankRAG with different retrievers.