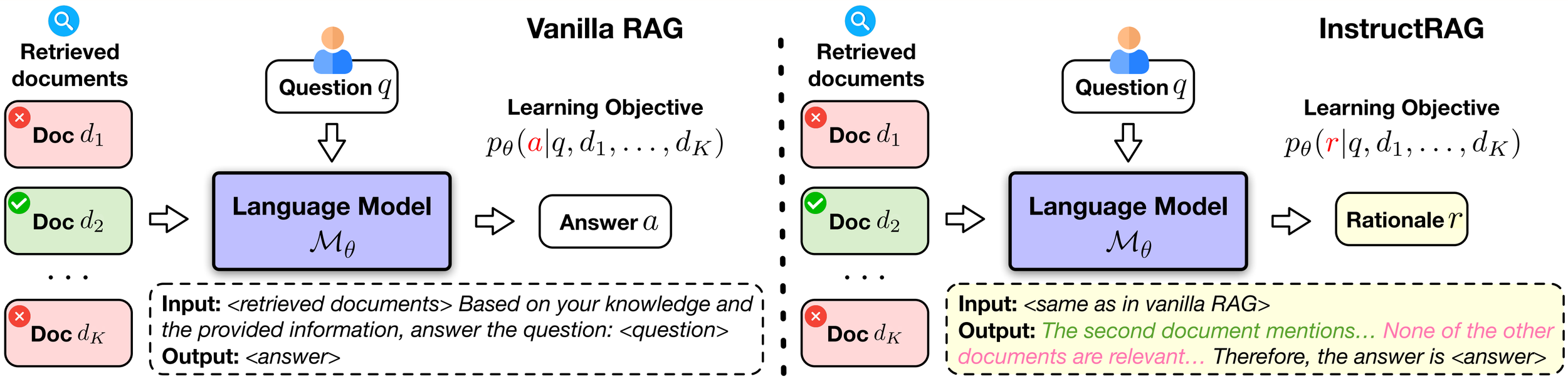
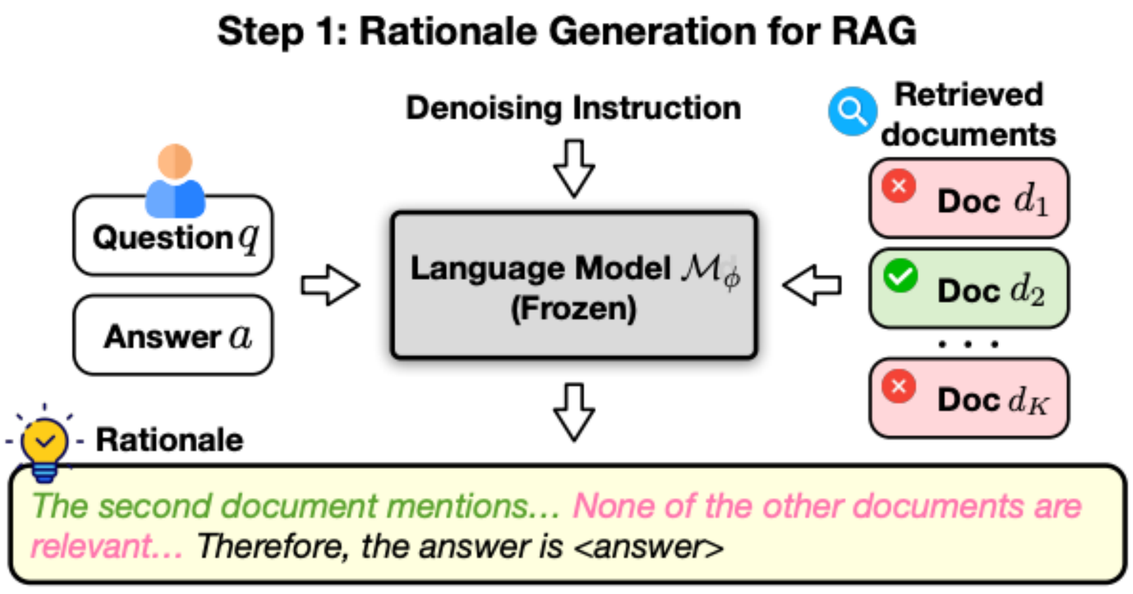
InstructRAG: Instructing Retrieval Augmented Generation via Self-Synthesized Rationales

Zhepei Wei, Wei-Lin Chen, Yu Meng
University of Virginia

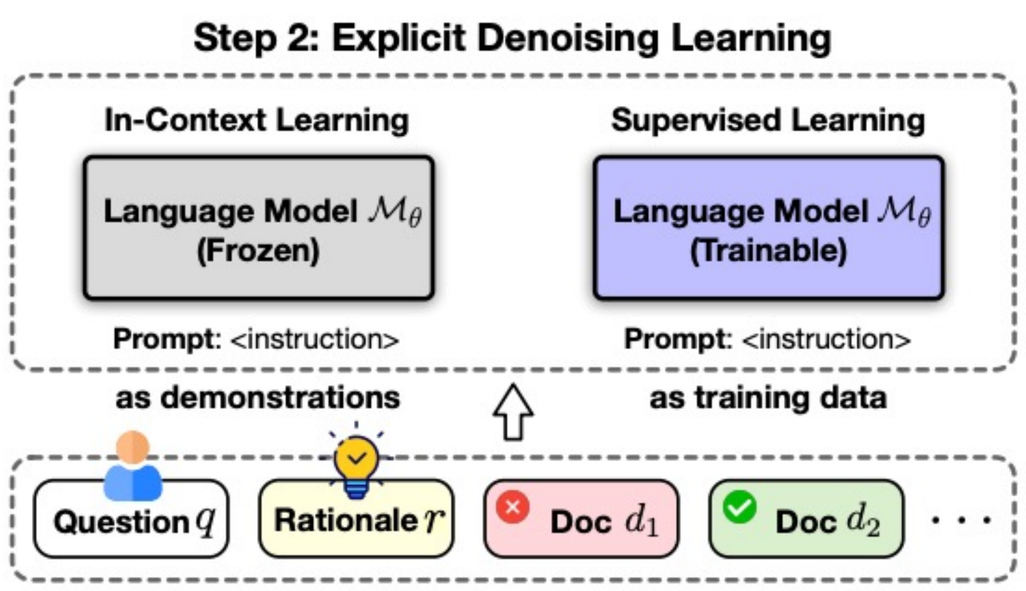
Vanilla RAG vs. InstructRAG



Step 1: Rationale Generation



Step 2: Denoising Learning



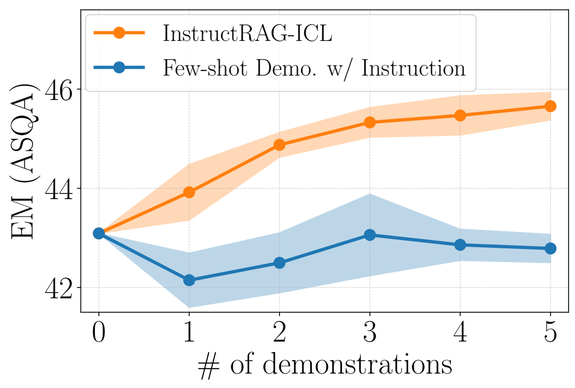
Evaluation & Analysis

Main Results

InstructRAG consistently outperforms baseline RAG methods across five knowledge-intensive benchmarks in both training-free and trainable settings.

Method	PopQA (acc)	TriviaQA (acc)	NQ (acc)	MultiHopQA (acc)	ASQA (em)	ASQA (pre)	ASQA (rec)
Baselines w/o Retrieval							
Vanilla Zero-shot Prompting							
ChatGPT*	29.3	74.3	–	–	35.3	–	–
Llama-3-Instruct$_{8B}$	22.8	69.4	46.6	45.6	30.6	–	–
Llama-3-Instruct$_{70B}$	28.9	80.6	57.9	57.5	39.1	–	–
RAG w/o Training							
In-Context RALM [70]							
ChatGPT*	50.8	65.7	–	–	40.7	65.1	76.6
Llama-3-Instruct$_{8B}$	62.3	71.4	56.8	43.4	40.0	62.1	66.4
Llama-3-Instruct$_{70B}$	63.8	76.3	60.2	51.2	43.1	62.9	67.6
Few-Shot Demo. w/ Instruction							
Llama-3-Instruct$_{8B}$	63.1	74.2	60.1	45.3	42.6	55.0	64.4
Llama-3-Instruct$_{70B}$	63.9	79.1	62.9	53.9	45.4	49.3	57.1
INSTRUCTRAG-ICL							
Llama-3-Instruct$_{8B}$	64.2	76.8	62.1	50.4	44.7	**70.9**	**74.1**
Llama-3-Instruct$_{70B}$	**65.5**	**81.2**	**66.5**	**57.3**	**47.8**	69.1	71.2
RAG w/ Training							
Vanilla Supervised Fine-tuning							
Llama-3-Instruct$_{8B}$	61.0	73.9	56.6	56.1	43.8	–	–
Self-RAG [3]							
Llama-2$_{7B}$	55.8	68.9	42.4	35.9	30.0	66.9	67.8
Llama-2$_{13B}$	56.3	70.4	46.4	36.0	31.4	**70.3**	**71.3**
Llama-3-Instruct$_{8B}$	55.8	71.4	42.8	32.9	36.9	69.7	69.7
RetRobust [105]							
Llama-2$_{13B}$	–	–	39.6	51.5	–	–	–
Llama-3-Instruct$_{8B}$	56.5	71.5	54.2	54.7	40.5	–	–
INSTRUCTRAG-FT							
Llama-3-Instruct$_{8B}$	**66.2**	**78.5**	**65.7**	**57.2**	**47.6**	65.7	70.5

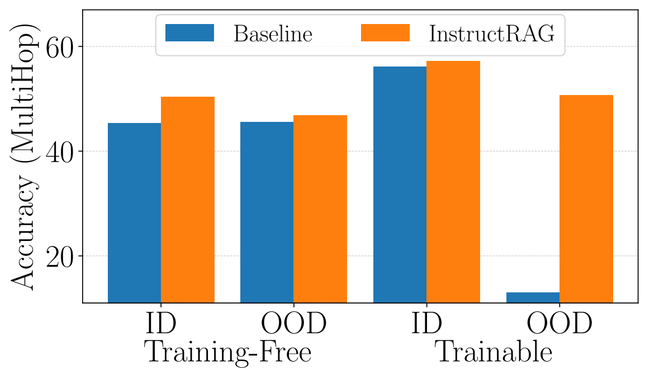
Demonstration Sensitivity



InstructRAG consistently benefits from more demonstrations.

Task Transferability

InstructRAG generalize well to unseen tasks.

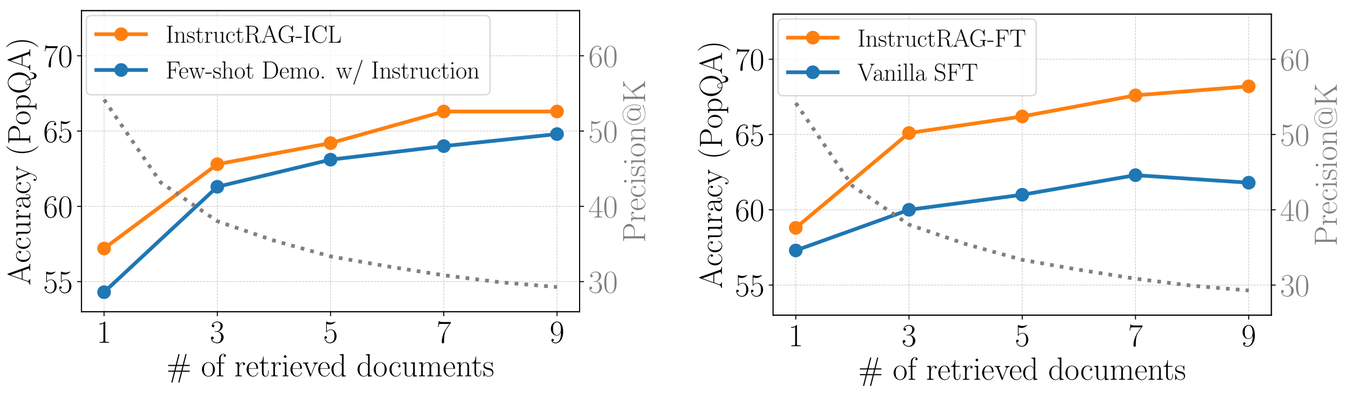


Method	pass@1	pass@10
Without Retrieval		
Llama-3-8B-Instruct	58.5	64.6
INSTRUCTRAG-FT	60.4	65.2
With Retrieval		
Llama-3-8B-Instruct	59.8	69.5
INSTRUCTRAG-FT	64.6	71.3

Single-hop QA (PopQA) → Multi-hop QA (2WikiMHQA)

QA Task (PopQA) → Non-QA Task (HumanEval)

Noise Robustness



InstructRAG is robust to increased noise ratios.

Evaluation with LLM-as-a-judge

Method	Pattern-based	LLM-based
RAG w/o Training		
In-Context RALM	56.8	64.5
INSTRUCTRAG-ICL	62.1	67.6
RAG w/ Training		
Vanilla SFT	56.6	65.1
INSTRUCTRAG-FT	65.7	69.7

GPT-4o-as-the-judge:
- Match semantic equivalence
- Lead to a more fair evaluation