RAS: Retrieval-And-Structuring for Knowledge-Intensive LLM Generation

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jiaming Shen✧, Jimeng Sun, Jiawei Han

University of Illinois Urbana-Champaign ✧Google DeepMind

paper code

Introduction

Motivation

Complex reasoning tasks demand comprehensive knowledge and structured thinking. LLMs struggle with knowledge-intensive reasoning:

- Lack of organized information for multi-step reasoning
- Hallucination from unstructured retrieved passages
- Failed implicit reasoning chains across context

Limitations of Existing Approaches

RAG: Single-pass retrieval misses critical facts; forces implicit logical bridging.

Static KGs: Costly ($85K–$500K at Wikipedia scale); blend conflicting evidence; introduce noise.

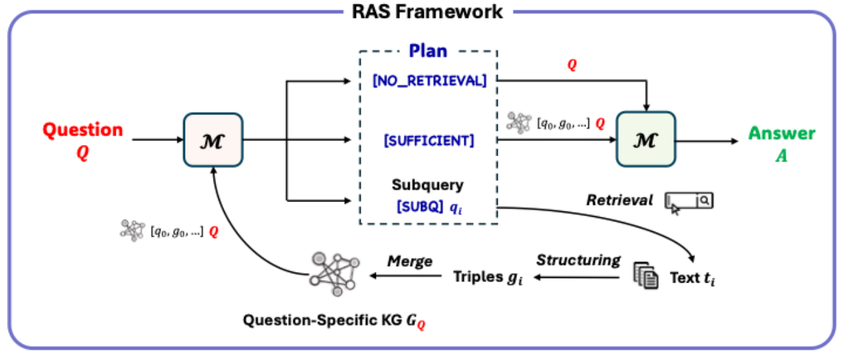
Performance Comparison

	Model/Method	Short-form TQA (acc)	Short-form 2WQA (F1)	Short-form PopQA (acc)	Closed-set Pub (acc)	Closed-set ARC (acc)	Long-form Generation ASQA (rouge)	ASQA (mauve)	ELI5 (rouge)	ELI5 (mauve)
Closed-source	**w/o Retrieval**									
	ChatGPT	74.3	24.8	29.3	70.1	75.3	36.2	68.8	22.8	32.6
	Sonnet-3.5	78.4	40.0	30.2	83.7	88.5	37.0	39.1	21.8	26.5
	w/ Single Retrieval (#docs=5)									
	Sonnet-3.5$_{\#docs=1}$	69.1	41.9	51.5	49.1	88.6	n/a	n/a	n/a	n/a
	Sonnet-3.5$_{\#docs=5}$	72.5	53.7	57.3	53.9	87.1	38.8	61.6	20.2	32.3
	SuRe$_{GPT-4o}$(Kim et al., 2024b)	72.3	38.1	53.6	57.2	79.6	36.0	74.2	19.2	51.6
	SuRe$_{Sonnet-3.5}$	76.8	37.6	41.2	62.8	91.6	30.2	69.9	15.4	27.2
	w/ Self-Reflective Retrieval									
	ReAct$_{Sonnet-3.5}$ (Yao et al., 2023)	73.4	53.7	55.0	62.2	89.2	38.8	61.6	20.2	32.3
	IRCoT$_{Sonnet-3.5}$ (Trivedi et al., 2023)	74.7	54.9	53.2	59.4	92.0	38.8	61.6	20.2	32.3
	Retrieval-And-Structuring (ours)									
	RAS$_{Sonnet-3.5}$	77.6	57.7	62.3	71.3	93.9	39.1	70.5	23.3	37.7
Open-source	**w/o Retrieval**									
	Llama2$_{7B}$	30.5	18.9	14.7	34.2	21.8	15.3	19.0	18.3	32.4
	Llama2$_{13B}$	38.5	20.2	14.7	29.4	29.4	12.4	16.0	18.2	41.4
	Llama3$_{8B}$	56.1	21.2	26.7	33.2	42.2	17.6	25.0	18.2	39.7
	w/ Single Retrieval (#docs=5)									
	Llama2$_{7B}$	42.5	21.0	38.2	30.0	48.0	22.1	32.0	18.6	35.3
	Llama2$_{13B}$	47.0	31.2	45.7	30.2	26.0	20.5	24.7	18.6	42.3
	Llama3$_{8B}$	60.4	33.4	48.6	36.5	40.1	23.9	52.1	18.8	40.7
	SuRe$_{7B}$ (Kim et al., 2024b)	51.2	20.6	39.0	36.2	52.7	35.8	76.2	16.1	26.6
	w/ Self-Reflective Retrieval									
	Self-RAG$_{7B}$ (Asai et al., 2023)	66.4	25.1	54.9	72.4	67.3	35.7	74.3	17.9	35.6
	Self-RAG$_{13B}$	69.3	26.9	55.8	74.5	73.1	37.0	71.6	18.7	38.5
	RPG$_{7B}$ (Lyu et al., 2024b)	65.1	33.6	56.0	73.4	65.4	37.6	84.4	19.1	46.4
	ReAct$_{7B}$ (Yao et al., 2023)	64.0	25.0	42.7	52.4	59.0	22.1	32.0	18.6	35.3
	IRCoT$_{7B}$ (Trivedi et al., 2023)	61.5	27.6	44.3	59.6	61.6	22.1	32.0	18.6	35.3
	Retrieval-And-Structuring (ours)									
	RAS$_{7B}$	72.7	42.1	58.3	74.7	68.5	37.2	95.2	19.7	47.8
	RAS$_{8B}$	73.8	44.2	57.7	77.6	71.4	37.6	96.2	20.1	54.4

Key Findings:

- Up to **8.7%** gains with proprietary LLMs, **7.0%** with open-source LLMs
- RAS$_7$B: +9.7% short-form QA, +7.9% long-form vs. Self-RAG & RPG
- Single-hop retrieval hurts on TQA & PubHealth — on-demand retrieval is essential
- Only 5% data (10K) already surpasses previous SOTA on TQA, 2WQA, ELI5

RAS Framework Overview



RAS dynamically constructs **question-specific knowledge graphs** through iterative retrieval and structured knowledge building

§3.1 Knowledge-Aware Planning

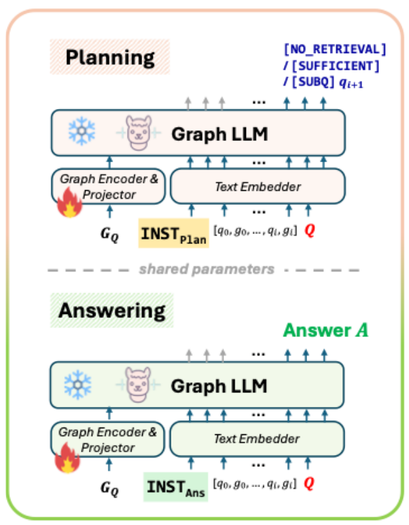
Dynamically assesses knowledge state; generates targeted sub-queries. Three actions: **[SUBQ]** (need more info) / **[SUFFICIENT]** (ready) / **[NO_RETRIEVAL]** (direct answer). Uses GNN-encoded evolving KG as input context.

§3.2 Text Retrieval & Structuring

Dense retrieval of top-k passages per sub-query. Lightweight text-to-triples model ft2t (LLaMA-3.2-3B) extracts (subject, predicate, object) facts. Sentence-BERT encodes node/edge attributes. Triples merged into evolving question-specific KG GQ.

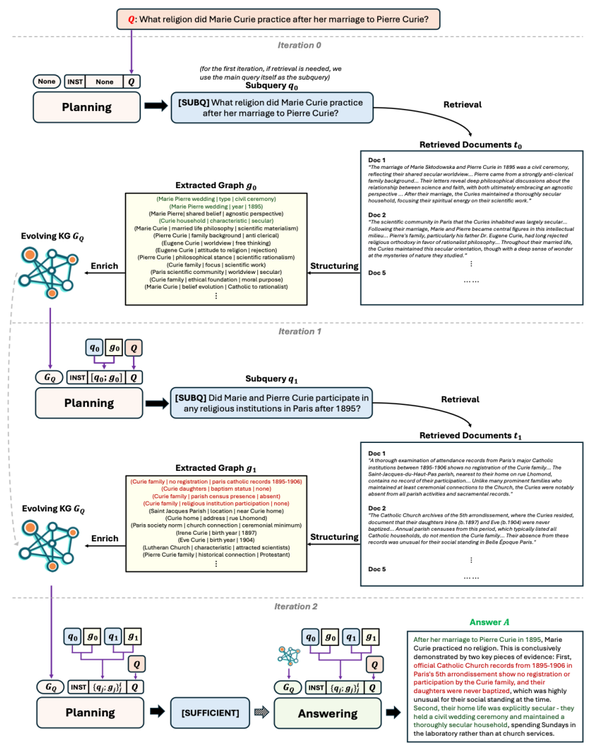
§3.3 Knowledge-Augmented Answering

Generates answer conditioned on accumulated KG and sub-query chain via Graph LLM architecture.



§3.4 Structure-Aware Multitask Training

Unified next-token prediction for planning + answering. LoRA fine-tuning with Graph Transformer encoder. Trained on HotpotQA-SUBQ (208K samples, derived from HotpotQA with Claude-3.5-Sonnet for doc filtering and sub-query generation).



Ablation & Analysis

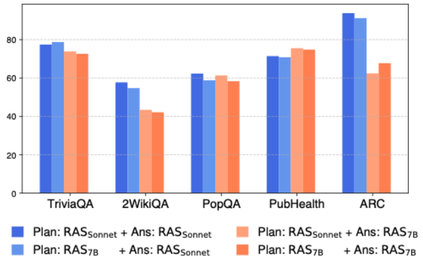
Ablation Highlights

→ Every component contributes; graph + LoRA + multitask = optimal

	TQA (acc)	2WQA (F1)	Pub (acc)	ASQA (rg)	ASQA (mv)
RAS$_{7B}$	**72.7**	**42.1**	**74.7**	**37.2**	**95.2**
Training Phase					
w/o GraphEncode	70.2	38.4	66.4	33.1	85.0
w/o LoRA	71.5	37.8	54.8	32.8	84.8
w/o Text-to-Triple	70.4	38.2	71.4	36.2	73.8
w/o Multi-Task	68.6	39.2	65.5	36.7	88.9
Inference Phase					
w/o Retrieval	56.9	27.4	69.0	31.3	70.6
w/o GraphEncode	68.8	38.7	67.3	36.5	93.6
w/o Planning	66.7	37.8	71.5	37.2	95.2

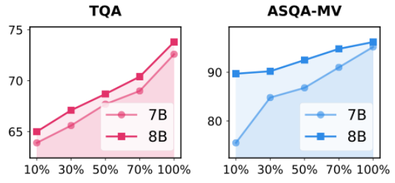
Role-Swapping

Performance limited by answering, not planning. RAS$_7$B planning ≈ Sonnet w/ 60× fewer params.



Graph Abundance

More triples → better performance; even 30–50% triples yield substantial gains. 8B > 7B



Impact of Training Volume

Performance scales with increasing training data