HydraRAG: Structured Cross-Source Enhanced LLM Reasoning

Xingyu Tan[1,2], Xiaoyang Wang[1], Qing Liu[2], Xiwei Xu[2], Xin Yuan[2], Liming Zhu[2], Wenjie Zhang[1]

[1] University of New South Wales, [2] Data61, CSIRO Email: xingyu.tan@unsw.edu.au

Introduction & Motivation

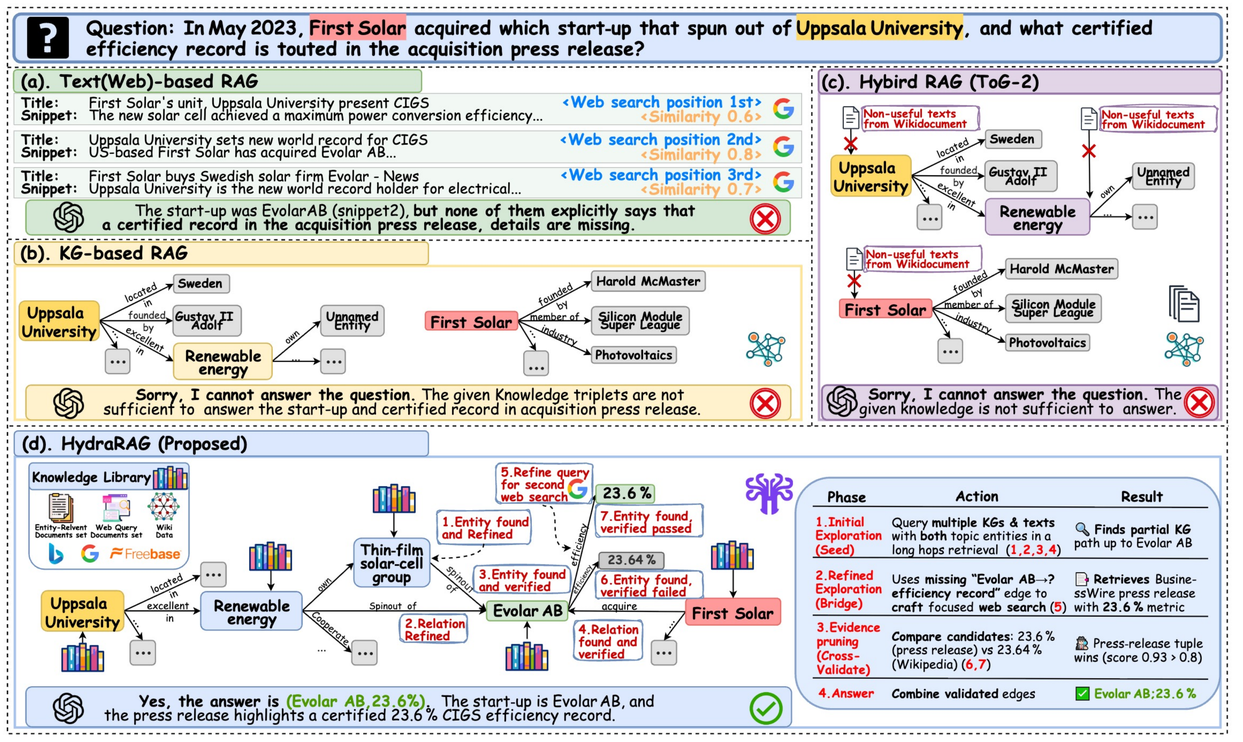
❑ **LLM Challenges:** LLMs struggle with complex reasoning, staying updated with current knowledge, and hallucination. Existing plan-retrieval-answering methods rely heavily on LLM reasoning rather than knowledge faithfulness.

❑ **Hybrid RAG & Motivation** : Current **Hybrid RAG system** retrieves evidence from **both knowledge graphs (KGs)** and **text documents** to support LLM reasoning.

- **Multi-source verification:** Existing methods rely on LLM semantics to merge evidence from different sources without assessing reliability or consistency.
- **Multi-hop reasoning:** They perform only local one-hop retrievals, which often miss global multi-hop paths needed for complete reasoning.
- **Multi-entity questions:** They treat each entity separately, generating redundant and noisy candidate paths that lower precision.
- **Graph structure utilization:** They fail to build unified graphs from multiple sources, preventing effective graph-based reasoning and pruning

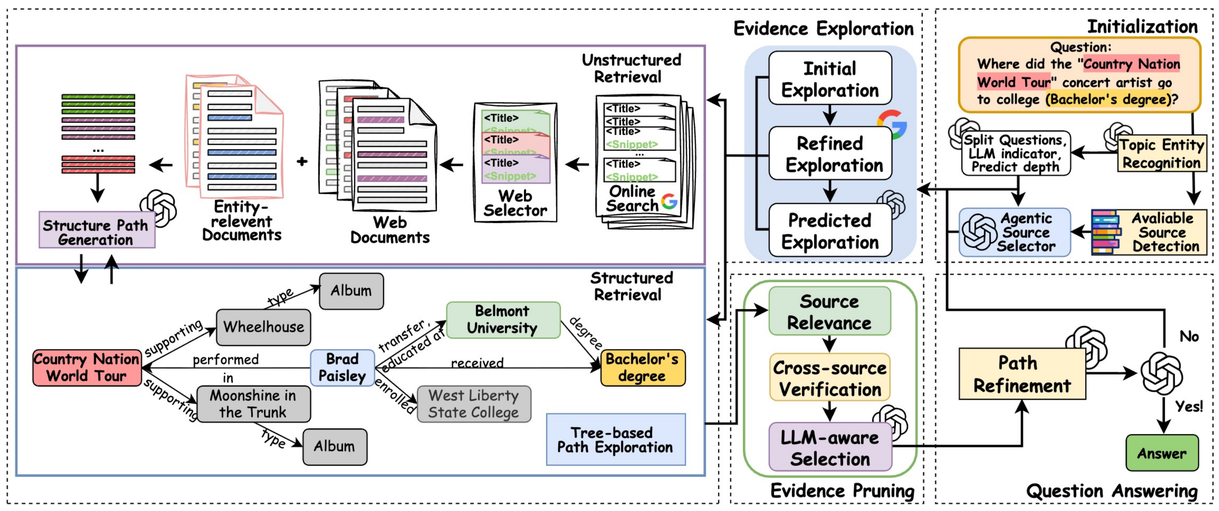
❑ **Our Contribution:**

- **Structured source-aware retrieval:** integrates heterogeneous evidence from diverse sources into a unified structured representation, enabling seamless reasoning.
- **Multi-source verification:** prunes candidate paths using both question relevance and cross-source corroboration, reducing hallucination and LLM cost.
- **Interpretable cross-source reasoning:** HydraRAG provides transparent, traceable reasoning paths showing how multi-modal facts lead to the answer.
- **Efficiency and adaptability:** a) HydraRAG is a **plug-and-play framework** that can be seamlessly applied to various LLMs, KGs, and texts. b) HydraRAG is **auto-refresh.** New information is incorporated instantly via web retrieval instead of costly LLM fine-tuning. c) HydraRAG **achieves SOTA** results on all the tested datasets, and enables smaller models to achieve reasoning performance comparable to GPT-4-Turbo.



Method

An example workflow of HydraRAG:

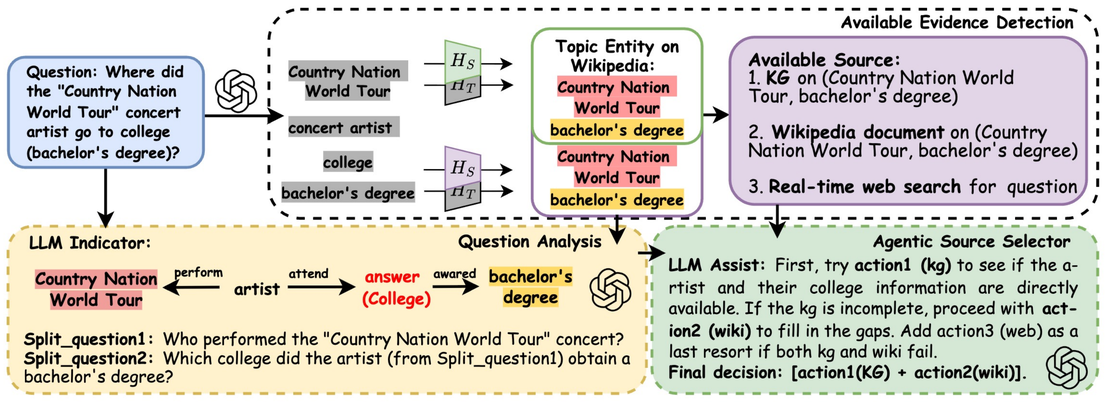


❖ Initialization

❑ Build available source detector for **facts evidence implements (LLM Lack Knowledge)**

❑ Skyline LLM indictor for **long reasoning guiding and length prediction (Multi-hops Problem)**

❑ **Agentic source selector (LLM Costs Reduction)**



❖ Exploration

❑ **Initial Exploration:**

- **Aim: Agentic selector-chosen sources** for high-precision seeds, **provide multi-hops and**
- **multi-entities faithful and interpretable facts** for reasoning
- **Dynamic exploration:** begins exploration from a predicted depth
- **Topic entity path :** Paths contained all the topic entity and meet the length's limitation considered.

❑ **Refined Exploration:**

- **Aim:** Leverage the **LLM sematic understanding** to **generate (refine) new query**, and **live web to patch gaps;**

❑ **Predicted Exploration:**

- **Aim:** Leverage the **inherent knowledge of the LLM** to generate predictive insights, and live web to patch gaps;
- generate one predict insight, then employ a **verification process**

❖Path Prune

❑ **Pruning = Cross-Score + LLM Selector**

❑ **Cross-Score = Source relevance + Cross-source verification**

- **Source relevance:** Given a query **skyline indicator** and its **topic-entity** set Topic(q), we compute a hybrid relevance score:
- **Cross-source verification:** We estimate the reliability of each candidate path using three reliability features:
  - **Source reliability**
  - **Corroboration from independent sources**
  - **Consistency with existing KG facts.**

❖Question Answering

❑ **Path Refinement:**

- Creating a **concise and focused** path
- To **decrease hallucinations caused by paths with excessive or incorrect text**

❑ **Question Answering:** encouraging deep reasoning

- **Deep reasoning:** prompt the LLM to answer individual split questions and then the overall question
- **Slow thinking:** use Chain-of-Thought to promote thorough thinking

Experiments & Results

❖ Experimental Result on Knowledge-Intensive Datasets

Table 1: Results of HydraRAG across all datasets, compared with the state-of-the-art (SOTA) with GPT-3.5-Turbo. The highest scores are highlighted in bold, while the second-best results are underlined for each dataset.

Type	Method	LLM	Multi-Hop KBQA				Single-Hop KBQA	Slot Filling	Open-Domain QA
			CWQ	WebQSP	AdvHotpotQA	QALD10-en	SimpleQA	ZeroShot RE	WebQuestions
LLM-only	IO prompt	GPT-3.5-Turbo	37.6	63.3	23.1	42.0	20.0	27.7	48.7
	CoT (Wei et al., 2022)	GPT-3.5-Turbo	38.8	62.2	30.8	42.9	20.3	28.8	48.5
	SC (Wang et al., 2023)	GPT-3.5-Turbo	45.4	61.1	34.4	45.3	18.9	45.4	50.3
Vanilla RAG	Web-based	GPT-3.5-Turbo	41.2	56.8	28.9	36.0	26.9	62.2	46.8
	Text-based	GPT-3.5-Turbo	33.8	67.9	23.7	42.4	21.4	29.5	35.8
KG-based RAG	ToG (Sun et al., 2024)	GPT-3.5-Turbo	58.9	76.2	26.3	50.2	53.6	88.0	54.5
	ToG (Sun et al., 2024)	GPT-4	69.5	82.6	-	54.7	66.7	88.3	57.9
	PoG (Tan et al., 2025)	GPT-3.5-Turbo	74.7	93.9	-	-	80.8	-	81.8
Hybrid RAG	CoK (Li et al., 2024c)	GPT-3.5-Turbo	-	77.6	35.4	47.1	-	75.5	-
	ToG-2 (Ma et al., 2025b)	GPT-3.5-Turbo	-	81.1	42.9	54.1	-	91.0	-
Proposed	HydraRAG-E	Llama-3.1-70B	71.3	89.7	48.4	70.9	80.4	95.6	76.8
	HydraRAG	Llama-3.1-70B	75.6	93.0	55.2	76.0	85.9	94.2	81.4
	HydraRAG-E	GPT-3.5-Turbo	76.8	94.0	51.3	81.1	81.7	96.9	85.2
	HydraRAG	GPT-3.5-Turbo	**81.2**	**96.1**	**58.9**	**84.2**	**88.8**	**97.7**	**88.3**

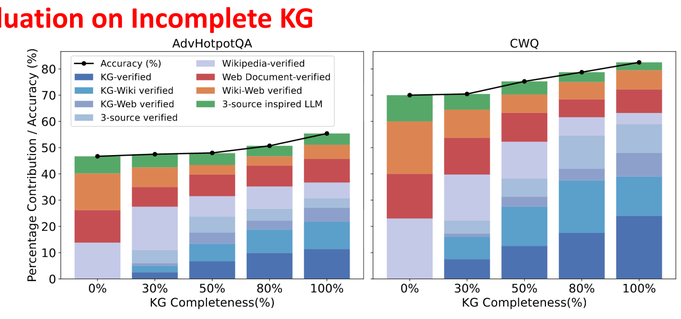
❖LLM Backbones Ablation

Table 2: Performance of the IO baseline and HydraRAG across four datasets on different backbone models. The highest improvement is highlighted in bold, while the second-best results are underlined for each model.

Dataset	Llama-3.1-8B			Llama-3.1-70B			DeepSeek-v3			GPT-3.5-Turbo			GPT-4-Turbo		
	IO	HydraRAG	%↑	IO	HydraRAG	%↑	IO	HydraRAG	%↑	IO	HydraRAG	%↑	IO	HydraRAG	%↑
AdvHotpotQA	16.9	35.6	111	21.7	48.4	123	27.8	55.4	99.0	23.1	56.2	143	46.4	67.9	46.0
WebQSP	38.5	86.0	123	56.2	95.2	69.0	68.0	97.7	44.0	66.3	96.9	46.0	75.4	98.2	30.0
CWQ	29.8	62.4	109	35.4	83.2	135	38.7	84.5	118	39.2	84.0	114	45.3	89.7	**98.0**
ZeroShot RE	27.2	77.5	**185**	34.6	97.5	**182**	38.6	97.0	**151**	37.2	97.7	**163**	49.8	98.5	**98.0**

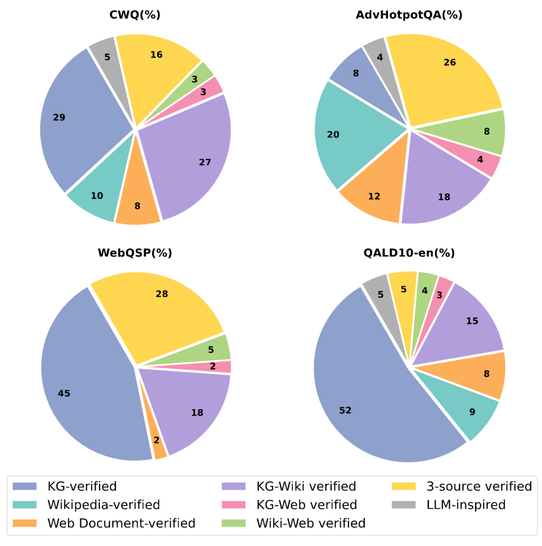
❑ **HydraRAG enables smaller models** (Llama-8B) to **achieve reasoning performance comparable to GPT-4-Turbo.**

❖Effective evaluation on Incomplete KG



❑ **HydraRAG does not solely depend on KG** data and **effectively mitigates KG incompleteness issues**, highlighting its adaptability.

❖Effective evaluation on Multi-Source Verification



Scan For HydraRAG

❑ Up to 56% of correct answers are **jointly verified by at least two distinct knowledge sources.** This demonstrates the strength of HydraRAG in **leveraging multi-source evidence, which is essential for faithful and interpretable reasoning**.