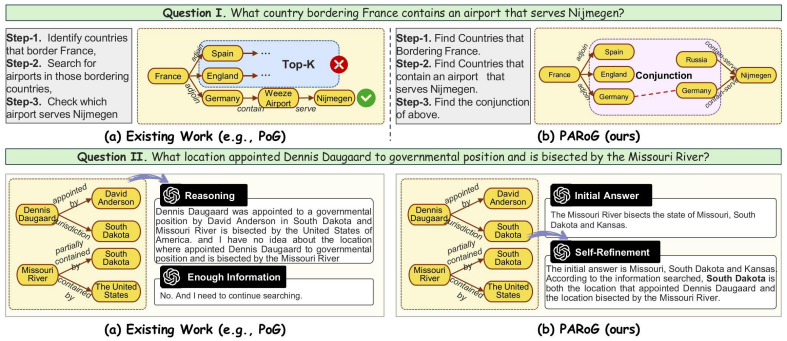
Plan-Answer-Refine-on-Graph: Structured Planning and Self-Refinement for Large Language Model Reasoning on Knowledge Graphs

Yuxin Shi, Han Fu, Zhuo Li, Chenghao Liu, Xiaoxue Ren, Jianling Sun

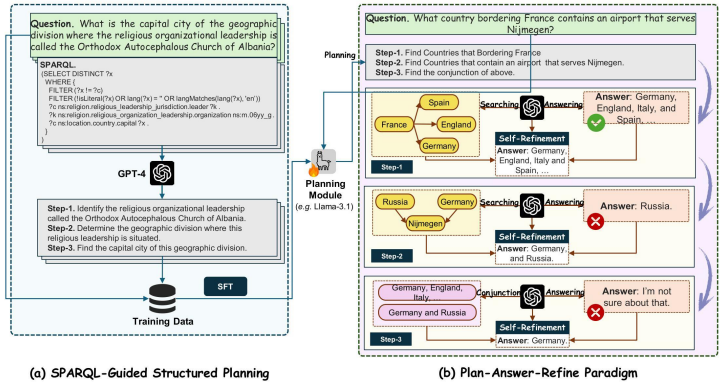
ICLR

INTRODUCTION



Existing methods fail because of truncated search paths and error amplification from faulty entities.

METHODS



PARoG combines structured planning with iterative answer refinement over retrieved KG evidence.

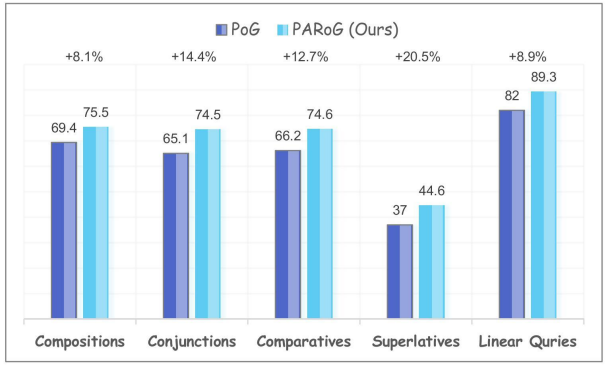
RESULTS

Methods	WebQSP	CWQ
LLM Prompting		
IO (Brown et al., 2020)	63.3	37.6
CoT (Wei et al., 2022)	62.2	38.8
SC (Wang et al., 2023c)	61.1	45.4
Graph-Retrieval Methods		
GNN-Rag (Mavromatis & Karypis, 2025)	82.8	62.8
SubgraphRag + GPT4o (Li et al., 2025)	87.1	54.9
LLM ⊗ KG with GPT-3.5		
ToG (Sun et al., 2024)	76.2	57.1
RoG (Luo et al., 2024)	81.5	52.6
KG-Agent (Jiang et al., 2025)	79.2	56.1
StructGPT (Jiang et al., 2023a)	75.2	55.2
PoG (Chen et al., 2024b)	82.0	63.2
ReKnowS (Wang et al., 2025)	81.1	58.5
PARoG	**89.0** (± 1.3)	**73.1** (± 0.9)
LLM ⊗ KG with GPT-4		
ToG (Sun et al., 2024)	80.7	65.4
KG-Agent (Jiang et al., 2025)	81.2	67.0
StructGPT (Jiang et al., 2023a)	79.5	64.7
PoG (Chen et al., 2024b)	87.3	75.0
ReKnowS (Wang et al., 2025)	83.8	66.8
PARoG	**91.2** (± 0.9)	**79.3** (± 1.1)

Method	# Para	WebQSP	GrailQA	CWQ
Ours	8B	**89.3**	**82.7**	**73.3**
GPT-3.5	~ 20B	83.2	76.9	65.2
Deepseek-R1	671B	88.5	80.2	68.7

GPT-4 backend. Under GPT-3.5, gains on WebQSP and CWQ further increase to +7.3 and +10.1 points.
8B planner model outperforms gpt-3.5 and deepseek-r1

different query types



Research question

- Existing LLM KG systems still struggle on compositional KGQA.
- Linear entity–relation walks create search-space truncation bias.
- One-shot retrieve-and-answer amplifies errors from faulty entities.

Approach

- Build structured plans from question–SPARQL pairs in WebQSP, CWQ, and GrailQA.
- Train a Llama-3.1-8B planner on 74,802 decompositions.
- For each sub-query: plan → answer → retrieve → refine
- Supports conjunctions, compositions, comparatives, and superlatives.

CONCLUSIONS

- Biggest gains appear on complex logical query types.
- Self-refinement corrects 62–77% of initially wrong answers.
- A SPARQL-supervised 8B planner can beat much larger generic planners.