ACL 2025 VIENNA

HELIOS: Harmonizing Early Fusion, Late Fusion, and LLM Reasoning for Multi-Granular Table-Text Retrieval

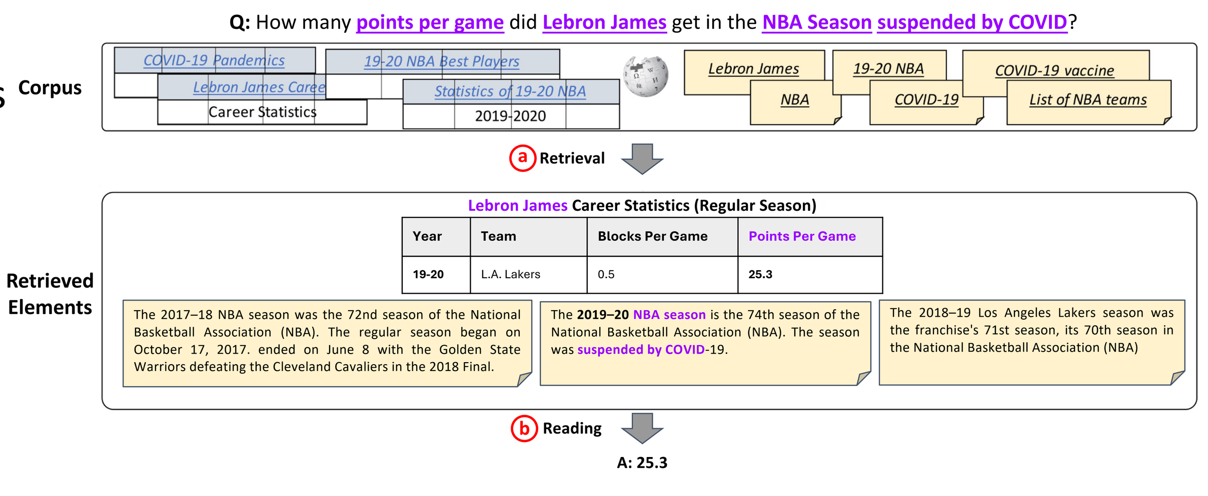
Sungho Park (POSTECH)
Joohyung Yun (POSTECH)
Jongwuk Lee (SKKU)
Wook-Shin Han* (POSTECH)

POSTECH POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
SUNGKYUNKWAN UNIVERSITY

My LinkedIn!
Our paper!

Open Table and Text Question Answering

Goal: Generate answer to a question by extracting answer strings from retrieved elements from a fixed corpus, a set of passages and tables.



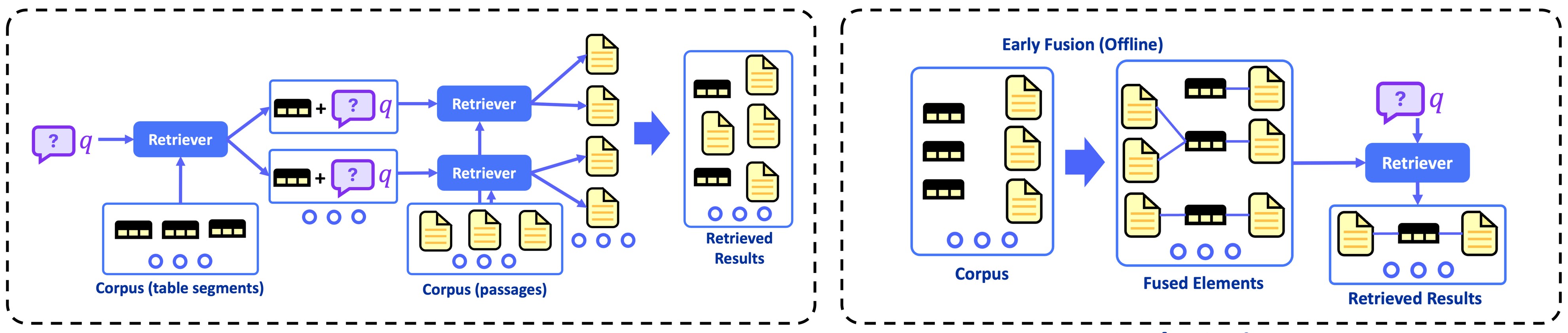
Standard OTT-QA System:

ⓐ Retrieval : Retrieve elements from a fixed corpus of passages and tables with a retriever.
ⓑ Reading : Analyze retrieved elements to provide an answer to the given question with a reader.

Previous Methods

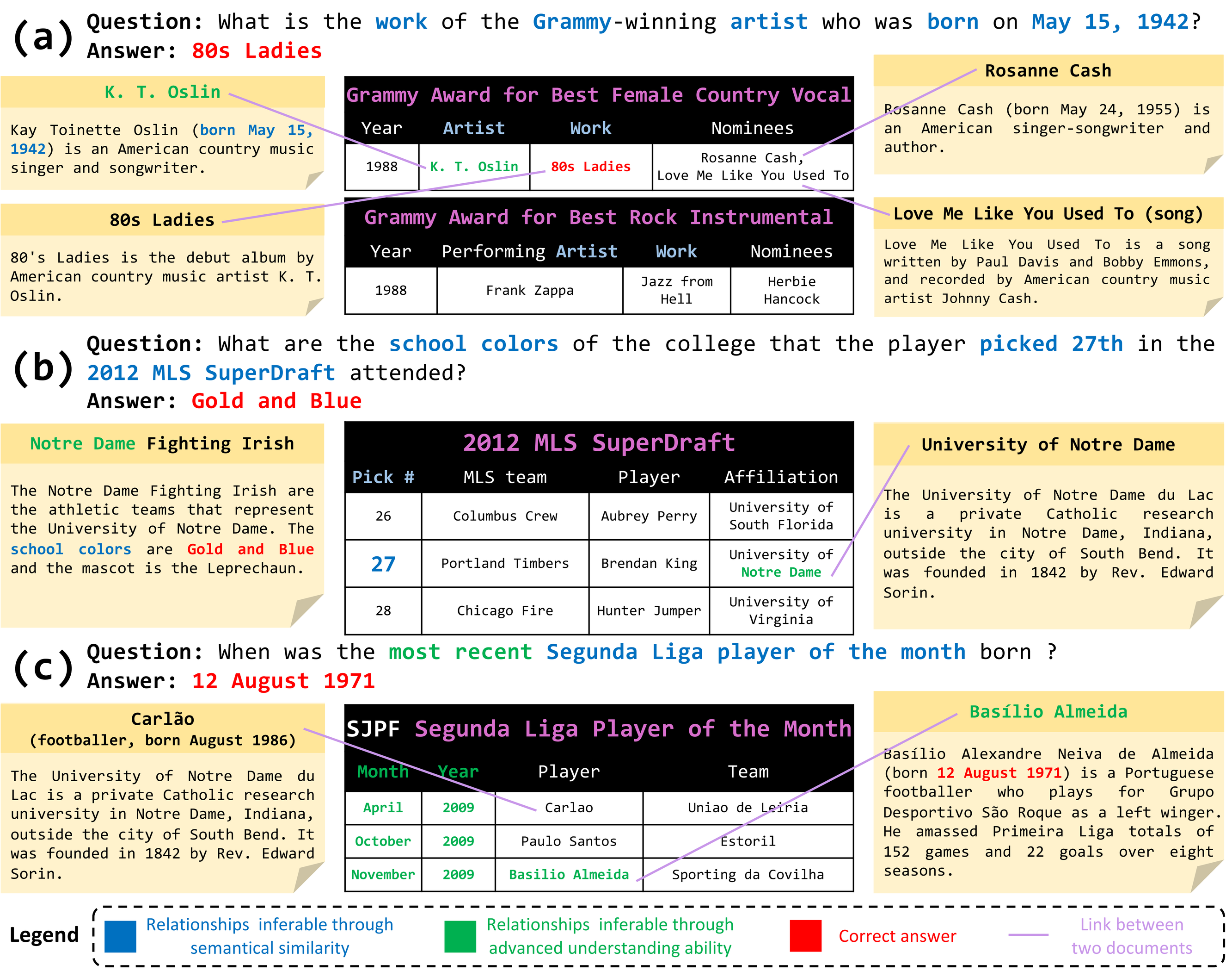
Existing studies are categorized into two main approaches based on when this relationship is considered:

- **Late fusion:** Relationship is considered after the query is provided.
- **Early fusion:** Relationship between tables and passages is considered before the query is provided.



Late Fusion — Early Fusion

What problem do we solve?



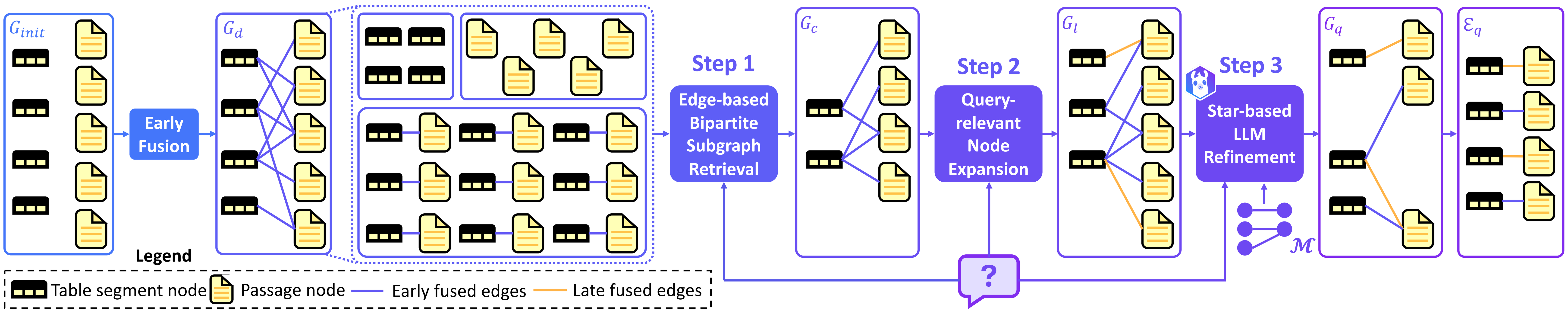
Retrieval methods for open-domain table-text QA still struggle with:

(a) overly coarse retrieval units
(b) missing query-specific links across tables and passages, and
(c) weak reasoning for multi-hop and aggregation questions.

How does HELIOS push the frontier?

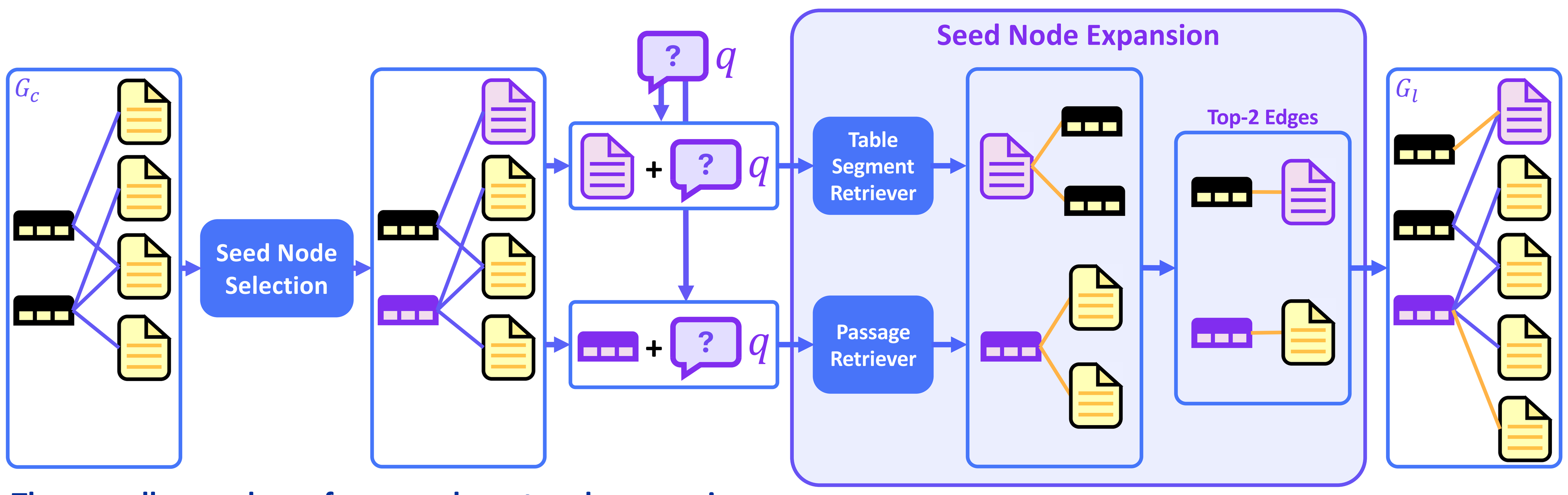
HELIOS **reframes retrieval as locating a query-relevant subgraph from a bipartite data graph** constructed via early fusion between table segments and passages. It introduces a three-stage, granularity-aware pipeline that **harmonizes** the strengths of both **early** and **late fusion** techniques while incorporating **LLM reasoning**.

System Overview



1. **Edge-based Bipartite Subgraph Retrieval** (Early Fusion): Offline entity linking builds the graph. An edge-level multi-vector retriever extracts a compact, high-precision candidate subgraph.
2. **Query-relevant Node Expansion** (Late Fusion): Identifies nodes most aligned with the query and selectively expands them, restoring essential query-dependent links.
3. **Star-based LLM Refinement** (LLM Reasoning): Decomposes the expanded graph into star-shaped subgraphs. An LLM performs aggregation and multi-hop reasoning, retaining only verified evidence.

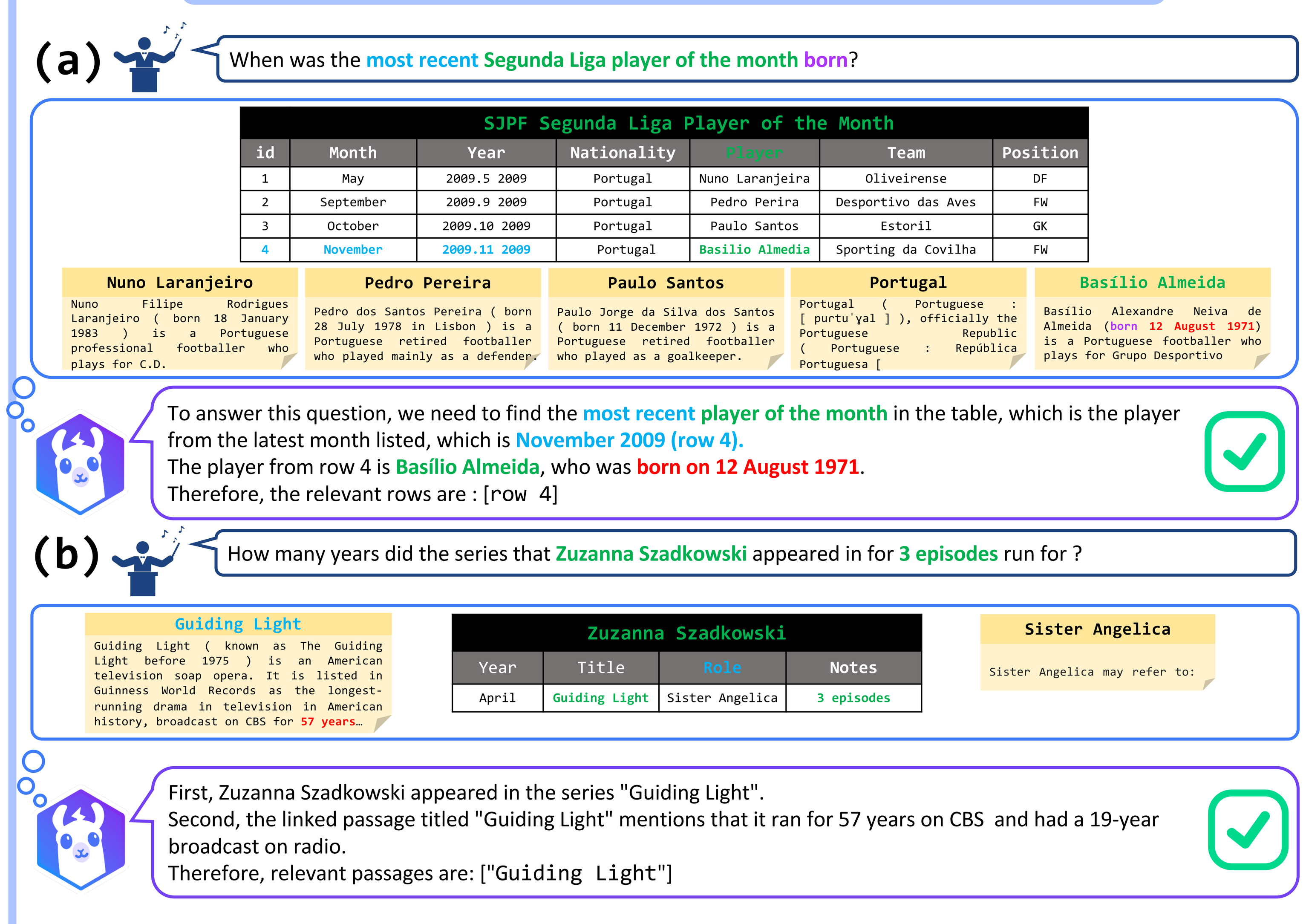
Query-relevant Node Expansion



The overall procedure of query-relevant node expansion

- The beam width b is set as 2 in this example.
- The purple-colored nodes indicate the selected seed nodes.

Star-based LLM Refinement



The star-based llm refinement consists of two phases: column-wise aggregation and passage verification.

- (a) shows a successful case of the column-wise aggregation module in resolving a complex query.
- (b) shows a successful case of the passage verification module in addressing the query.

Evaluation

* Please read our paper for full experiment results!

Retrieval Accuracy Comparison

Model	AR@2	AR@5	AR@10	AR@20	AR@50	nDCG@50	HITS@4k
Iterative Retriever	–	–	–	–	–	–	27.2
Fusion Retriever	–	–	–	–	–	–	52.4
OTTeR†	31.4	49.7	62.0	71.8	82.0	25.9	70.1
DoTTeR†	31.5	51.0	61.5	71.9	80.8	26.7	70.3
CORE†	35.3	50.7	63.1	74.5	83.1	25.4	77.2
COS†	44.4	61.6	70.8	79.5	87.8	33.6	81.8
COS w/ ColBERT & bge†	49.6	68.2	78.7	85.0	91.7	36.5	85.9
DoTTeR + COS + LLM†	50.0	62.4	70.0	76.2	84.7	34.7	–
HELIOS	**63.3**	**76.7**	**85.0**	**90.4**	**94.2**	**47.0**	**91.8**

HELIOS consistently outperforms all competitors on AR@k across different k values.

- upto **39.9% higher nDCG@50** compared to SOTA
- upto **12.2% higher HITS@4K** compared to SOTA

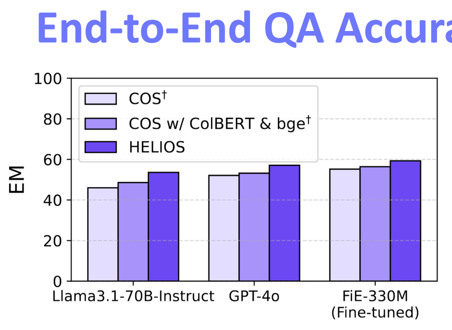
Ablation Study & Algorithm Execution Time

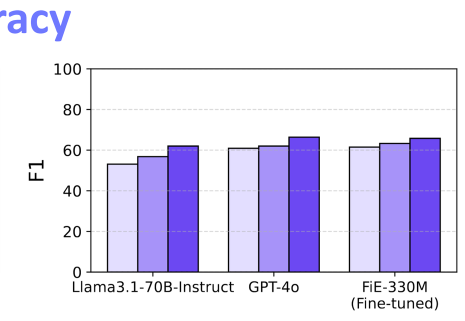
Algorithm	Execution Time (s)	nDCG@50
DoTTeR	0.08	26.7
CORE	4.13	25.4
COS	3.75	33.6
COS w/ ColBERT & bge	5.46	36.5
HELIOS	5.14	47.0
HELIOS w/ Finetuned SLR	4.76	**47.6**
w/o QNE	–	45.1
w/o SLR	2.16	46.5
w/o (SLR & Edge Reranker)	**1.11**	42.1

HELIOS achieved the best retrieval accuracy

- **reaching 47.6 nDCG@50** with finetuned SLR,
- while maintaining **comparable execution time**, and
- ablation confirms that both **SLR** and **QNE provide meaningful gains**.

End-to-End QA Accuracy





HELIOS consistently outperformed all reader models,

- achieving an average **EM gain of 7.5%**, and
- an average **F1 gain of 6.6%** over COS with ColBERT and BGE.

Conclusion

- We presented **HELIOS**, a novel table-text retrieval method that **harmonizes** the strengths of both **early** and **late fusion** techniques while incorporating **LLM reasoning**.
- It addresses the limitations of competitors by introducing a **multi-granular retrieval system** that optimally balances granularity across retrieval stages.
- Experiments on OTT-QA show that it **surpasses SOTA models**, achieving a 42.6% AR@2 improvement and a 39.9% nDCG@50 gain.