GEAR: Graph-enhanced Agent for Retrieval-augmented Generation

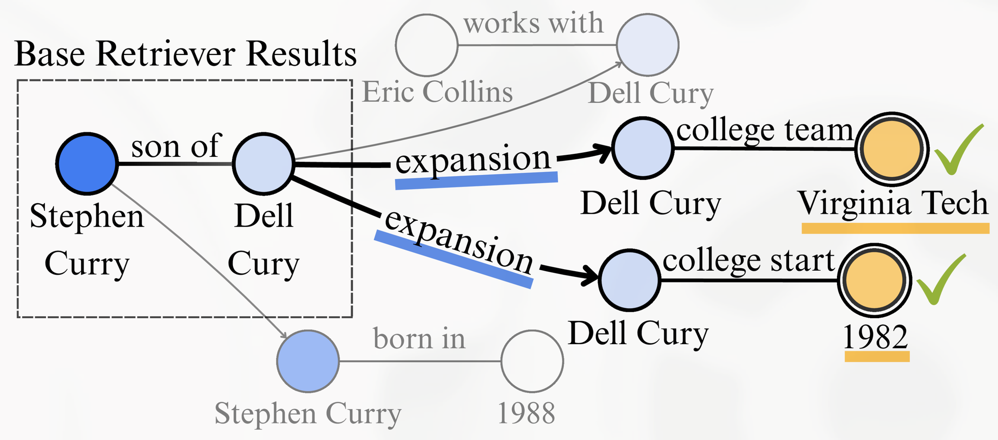
Zhili Shen† Chenxin Diao† Pavlos Vougiouklis† Pascual Merita†
Shriram Piramanayagam Enting Chen Damien Graux Andre Melo
Ruofei Lai Zeren Jiang Zhongyang Li Ye Qi Yang Ren Dandan Tu Jeff Z. Pan

Huawei Poisson Lab, UK https://gear-rag.github.io

ACL 2025 VIENNA JULY 27 - AUGUST 1

Multi-hop QA Example

"In what year did Stephen Curry's father join the team from which he started his college basketball career?"



- A base retriever cannot, by design, retrieve all necessary information in a single step
- Graph expansion enables retrieval of subsequent hops
- Guides the system toward the correct answer without using an LLM

[TL;DR] Why GEAR?

Task: Multi-hop Question & Answering

State-of-the-art retrieval performance: ≥ 17% relative ↑ for R@5 on MuSiQue vs HippoRAG w/ IRCoT
→ Leads to more than 35% relative ↑ for end-to-end QA F1

Less computational LLM workload: achieves best performance in less iterations, and sometimes w/o using multi-step agent

Neuroscience-inspired framework: modelling the communication between *hippocampus* and *neocortex* in the brain

A Walk-through

When did the location of the basilica which is named for the same saint that the Bremen Cathedral is named for become a country?

Offline Index Building Stage

For each passage $c_i \in \mathbf{C} = \{c_1, c_2, \ldots, c_C\}$, an LLM extracts a triple set, such that each triple is uniquely linked to one single passage.

1. Base Retrieval

For a query $\mathbf{q}$, $\mathbf{C}'_{\mathbf{q}} = h^k_{\text{base}}(\mathbf{q}, \mathbf{C})$ is a list of passages given by the retriever, implemented as BM25, SBERT, or a mix of both.

$\mathcal{P}_1$ Bremen Cathedral $\mathcal{P}_2$ Münster Cathedral
$\mathcal{P}_3$ Basilica of the Sacred Heart
$\mathcal{P}_4$ Saint Justin's Church, Frankfurt-Höchst
$\mathcal{P}_5$ Alatri Cathedral

2. Reader

An LLM `reads` $\mathbf{C}'_{\mathbf{q}}$ and summarises knowledge triples, outputting a collection $\mathbf{T}'_{\mathbf{q}}$ of triples: the *proximal triples*.

$\mathcal{T}'_1$ ⟨Bremen Cathedral, dedicated to, St. Peter⟩
$\mathcal{T}'_2$ ⟨Alatri Cathedral, dedicated to, Saint Paul⟩
$\mathcal{T}'_3$ ⟨Alatri Cathedral, co-cathedral of, Diocese Anagni-Alatri⟩
$\mathcal{T}'_4$ ⟨Bremen, is located in, Germany⟩

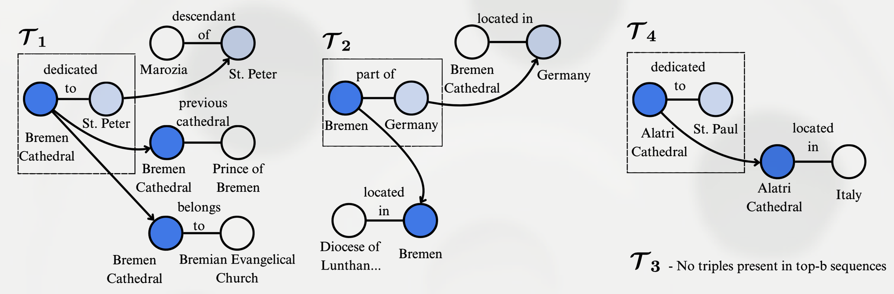
3. `tripleLink`

Initial nodes $\mathbf{T}_{\mathbf{q}}$ for graph expansion are identified by linking each triple in $\mathbf{T}'_{\mathbf{q}}$ to a triple in $\mathbf{T}$, using the `tripleLink` function.

$\mathcal{T}_1$ ⟨Bremen Cathedral, dedicated to, St. Peter⟩
$\mathcal{T}_2$ ⟨Alatri Cathedral, dedicated to, Saint Paul⟩
$\mathcal{T}_3$ ⟨Diocese of Macerata-Tolentino-Recanati-Cingoli-Treia, type, co-cathedral⟩
$\mathcal{T}_4$ ⟨Bremen, part of, Germany⟩

4. Graph Expansion

The primary component of graph expansion is *Diverse Triple Beam Search*. Here, we explore neighbourhood of a triple (defined as other triples with shared head or tail entities) and maintain top-b sequences (beams) of triples.



5. Gist Memory

Similar to the Reader, an LLM reads a collection of retrieved paragraphs $\mathbf{C}_{\mathbf{q}^{(n)}}$ and extracts an array of proximal triples $\mathbf{T}^{\mathcal{G}}_{\mathbf{q}^{(n)}}$, which are stored in the Gist Memory $\mathcal{G}^{(n)}$.

$\mathcal{T}_1$ ⟨Bremen Cathedral, dedicated to, St. Peter⟩
$\mathcal{T}_2$ ⟨Alatri Cathedral, dedicated to, Saint Paul⟩
$\mathcal{T}_3$ ⟨Lund Cathedral, dedicated to, Saint Lawrence⟩
$\mathcal{T}_4$ ⟨Bremen, part of, Germany⟩

6. Reasoner

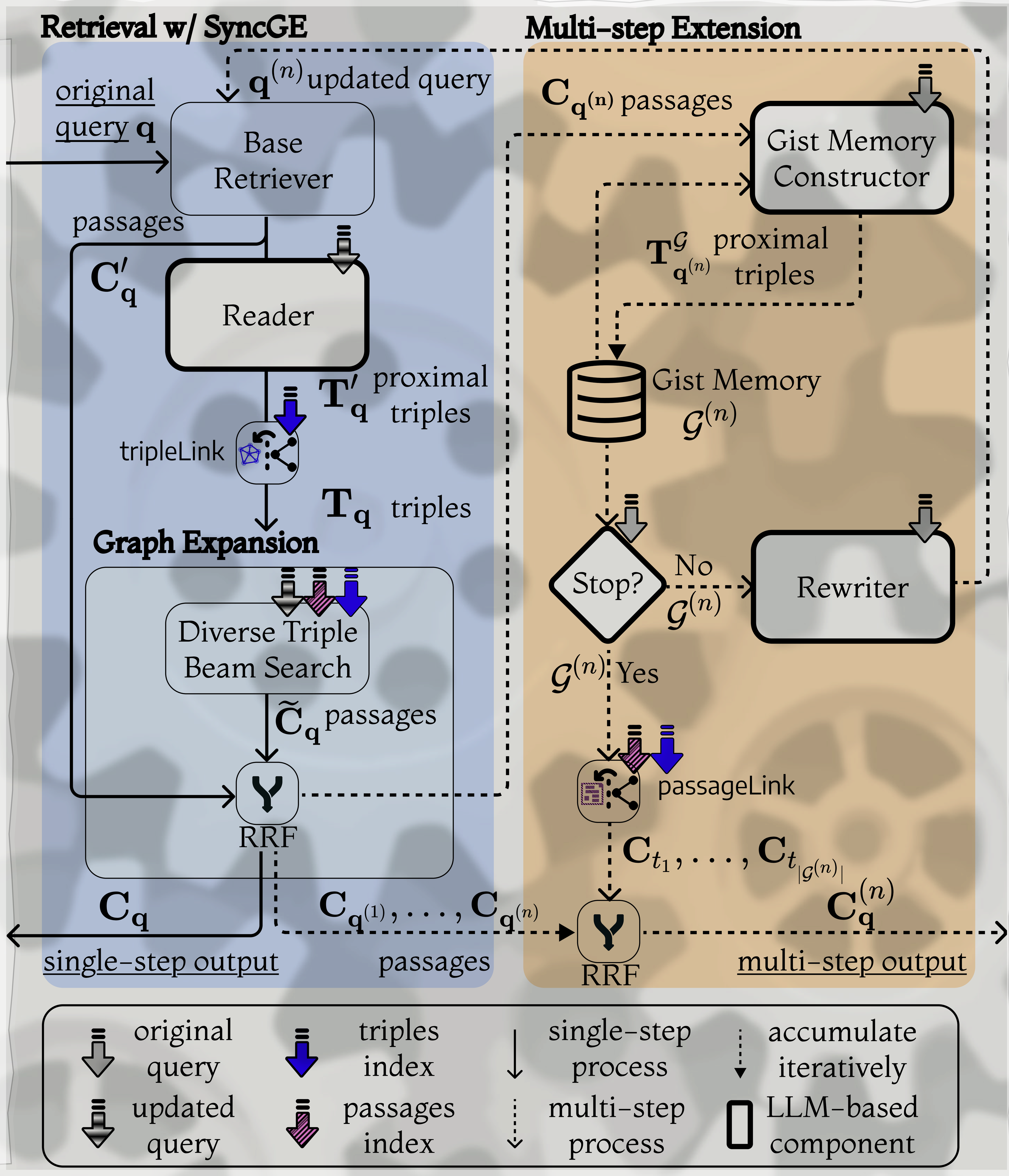
After updating $\mathcal{G}^{(n)}$, we assess whether it contains sufficient evidence to answer the original question via an LLM reasoning step.

Answerable: False **Answer or reason:** The provided facts do not contain information about the location of the basilica named for St. Peter, nor do they provide any details about when it became a country. The facts only mention the dedication of other cathedrals to different saints.

7. Rewriter

Given the original $\mathbf{q}$, the accumulated memory $\mathcal{G}^{(n)}$, and the reasoning output $\mathbf{r}^{(n)}$, an LLM is used to re-write the query.
We return to **step 1** and repeat.

Next query: What is the location of the basilica dedicated to St. Peter, and when did that location become a country?



Retriever	MuSiQue R@5	MuSiQue R@10	MuSiQue R@15	2Wiki R@5	2Wiki R@10	2Wiki R@15	HotpotQA R@5	HotpotQA R@10	HotpotQA R@15
ColBERTv2	39.4	44.8	47.7	59.1	64.3	66.2	79.3	87.1	90.1
HippoRAG	41.0	47.0	51.4	**75.1**	**83.2**	**86.4**	79.8	89.0	92.4
BM25	33.8	38.5	41.3	59.5	62.7	64.1	74.2	83.6	86.3
+ NaiveGE	37.5	45.5	48.4	65.0	70.7	71.8	79.1	89.1	91.9
+ SyncGE	44.7	52.6	57.4	70.5	76.1	79.3	87.4	93.0	94.0
SBERT	31.1	37.9	41.6	41.2	48.1	51.5	72.1	79.3	84.0
+ NaiveGE	32.2	41.4	45.4	45.1	54.0	57.3	76.1	84.7	88.8
+ SyncGE	41.6	51.3	54.2	54.8	64.9	70.7	84.1	89.6	92.8
Hybrid	39.9	46.3	49.1	60.0	65.8	66.6	77.8	85.8	89.7
+ NaiveGE	41.8	49.4	53.0	63.0	70.8	72.6	80.6	89.4	92.7
+ SyncGE	**48.7**	**57.7**	**61.2**	72.6	80.9	82.4	**87.4**	**93.3**	**95.2**
IRCoT (BM25)	46.1	54.9	57.9	67.9	75.5	76.1	87.0	92.6	92.9
IRCoT (ColBERTv2)	47.9	54.3	56.4	60.3	86.6	69.7	86.9	92.5	92.8
HippoRAG w/ IRCoT	48.8	54.5	58.9	82.9	90.6	93.0	90.1	94.7	95.9
GEAR	**58.4**	**67.6**	**71.5**	**89.1**	**95.3**	**95.9**	**93.4**	**96.8**	**97.3**

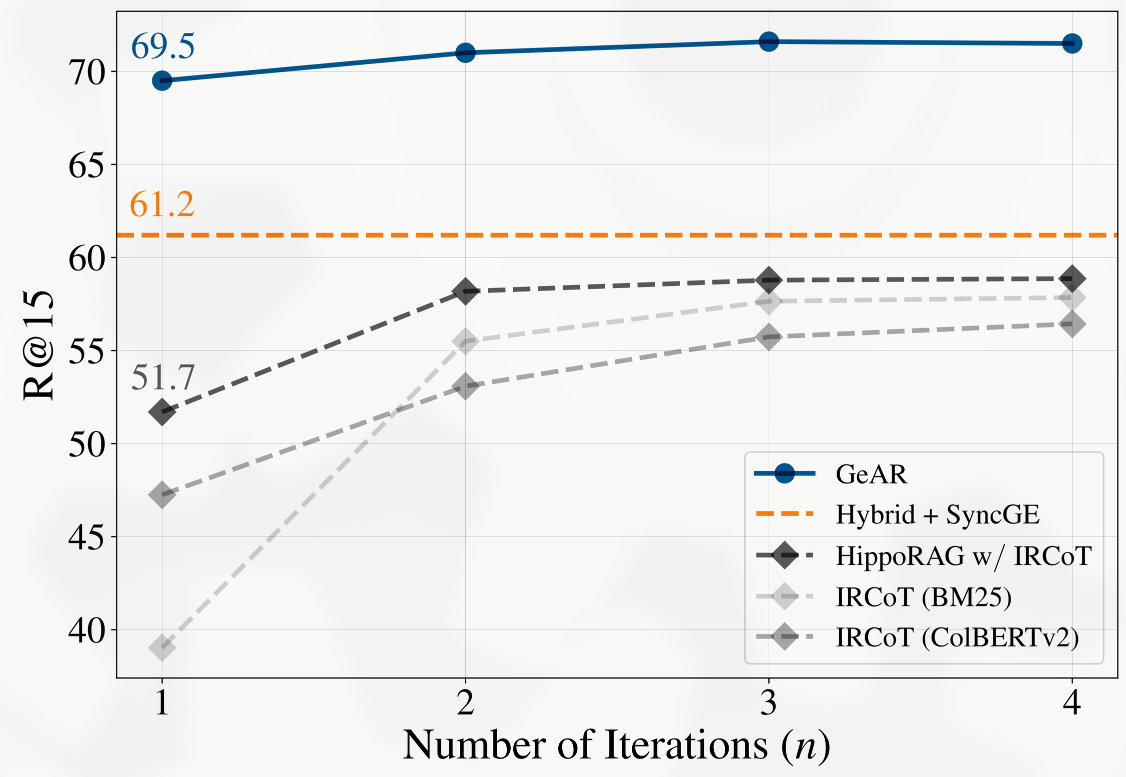
Metric	Dataset	w/ Diversity	w/o Diversity
R@5	MuSiQue	**48.7**	47.0
	2Wiki	**72.6**	68.2
	HotpotQA	**87.4**	85.0
R@10	MuSiQue	**57.7**	53.9
	2Wiki	**80.9**	76.0
	HotpotQA	**93.3**	92.2
R@15	MuSiQue	**61.2**	58.4
	2Wiki	**82.4**	77.4
	HotpotQA	**95.2**	94.3

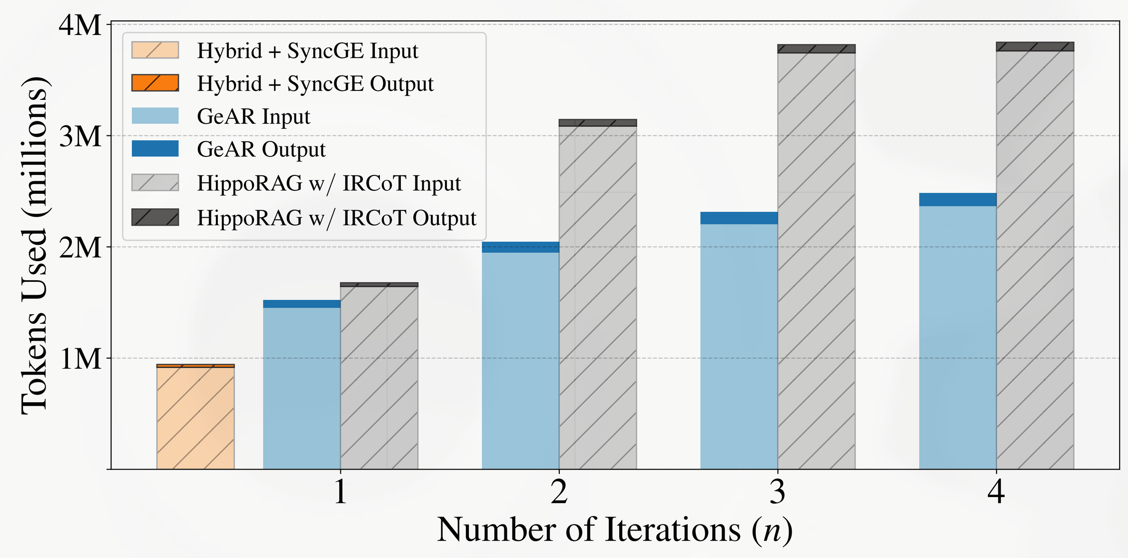
What makes GEAR work?

- [GE] Graph expansion on top of a base retriever
- [SyncGE] LLM for locating initial nodes for GE—synergetic behaviour between LLM and GE >> NaiveGE
- [Diversity Weights] Introducing diversity weight for triple beam search

Is GEAR efficient?

- GEAR requires fewer iterations than the competition to reach its maximum recall performance
- GEAR is more efficient in terms of LLM token utilisation
- Even for a single iteration, GEAR uses fewer tokens than HippoRAG w/ IRCoT, with substantially higher Recall@15





LaTeX TikZposter