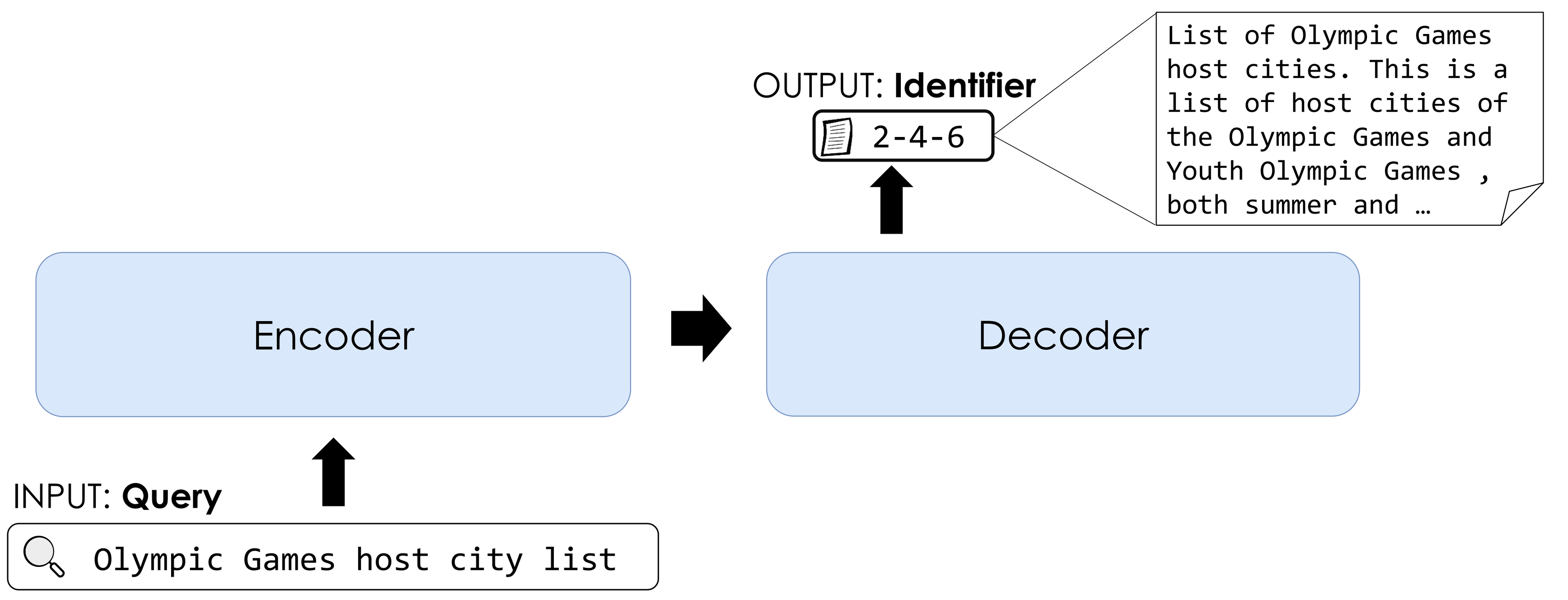
GLEN: Generative Retrieval via Lexical Index Learning

Sunkyung Lee*, **Minjin Choi***, **and Jongwuk Lee**
Sungkyunkwan University (SKKU), Republic of Korea
* Equal contribution

EMNLP 2023

Generative Document Retrieval

It aims to generate the **identifier of relevant documents** for a given query.
- All the corpus is encoded in model parameters, enabling end-to-end optimization.
- The index structure is unnecessary.



Takeaways

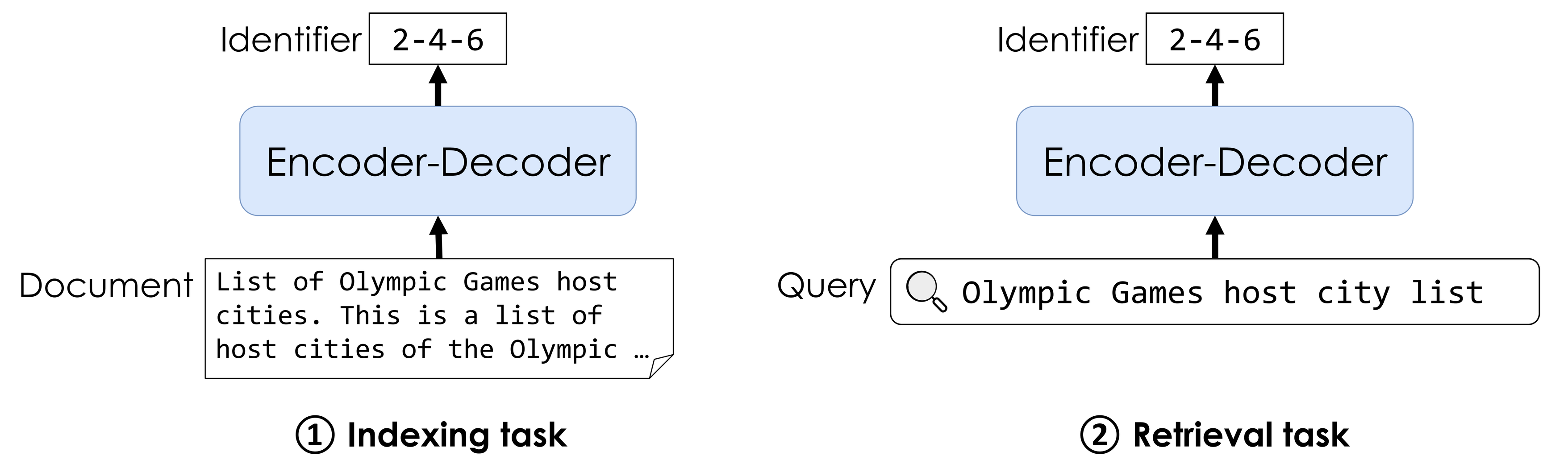
✓ A novel generative retrieval method for **dynamically learning lexical identifiers** based on query-document relevance
✓ **Two-stage lexical index learning** for dynamically learning lexical identifiers based on ranking signals
✓ **Collision-free inference** for efficient ranking using identifier weights

Motivation

How can we **learn** appropriate **document identifiers** for generative retrieval?

Limitation of Existing Methods

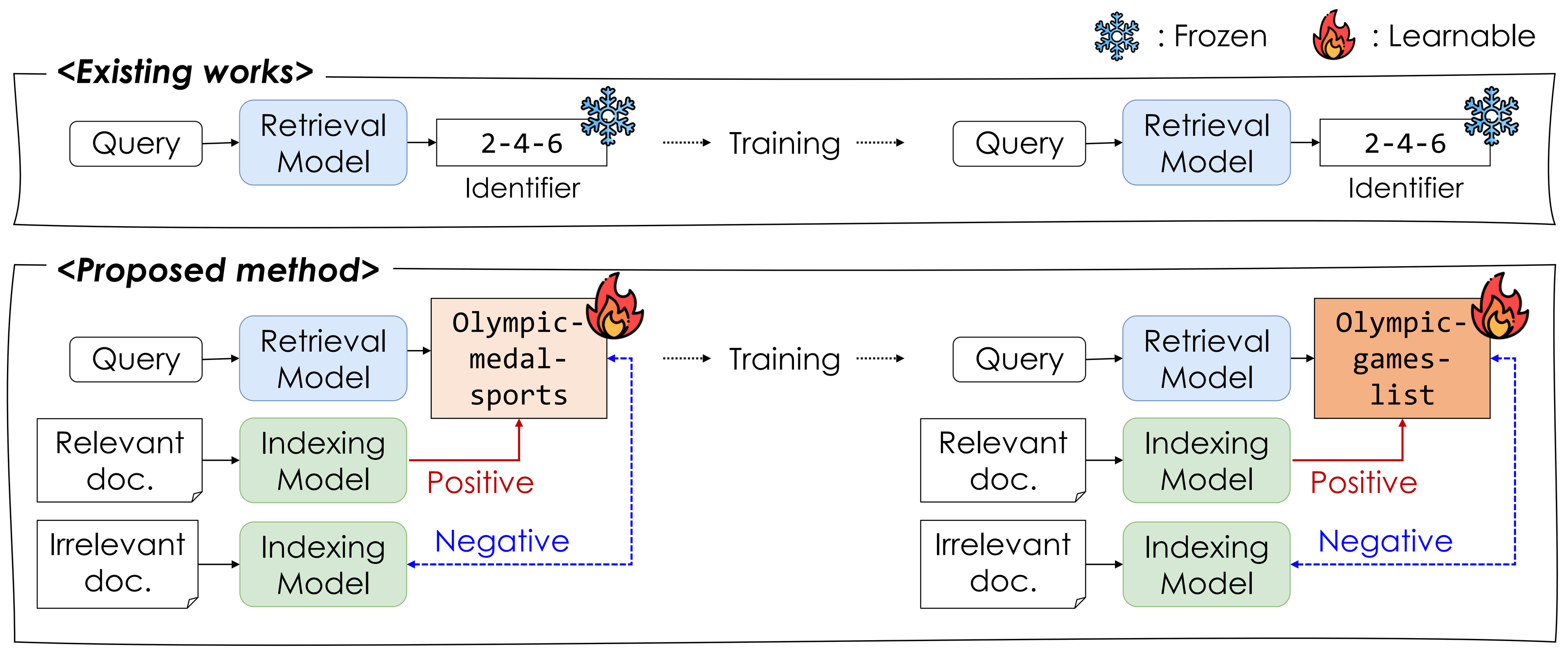
Most existing models **pre-define static document identifiers**, but they are **difficult to generalize** new documents.
- DSI (NeurIPS' 22) parameterizes a retrieval system with a single transformer.
- Static identifiers can be random numbers, topic hierarchies, titles, or URLs.



*Yi Tay et al. Transformer Memory as a Differentiable Search Index. NeurIPS 2022

Our Solution: Lexical Index Learning

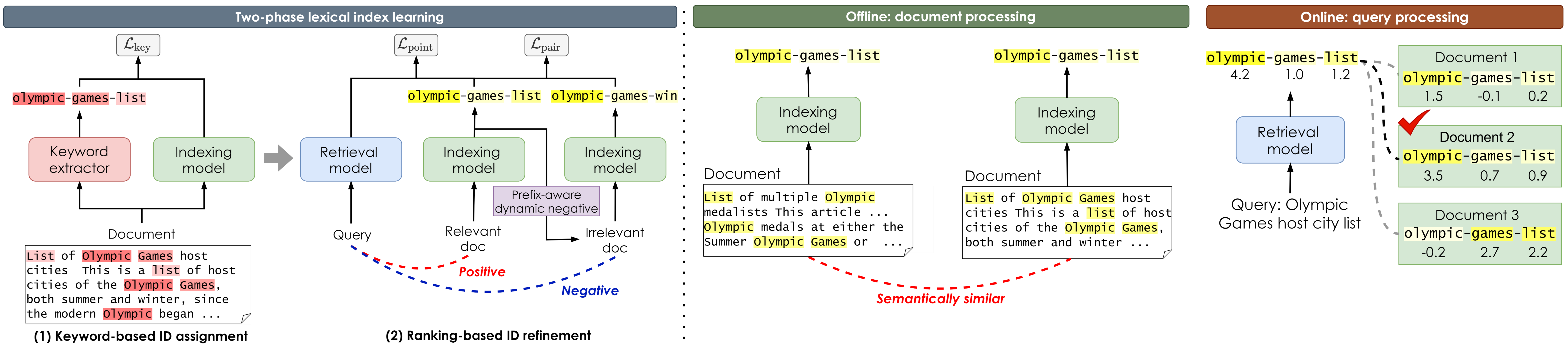
We propose a **lexical index learning** to dynamically learn identifiers considering **query-document relevance**.



GLEN: Generative Retrieval via Lexical Index Learning

For training, GLEN introduces a dynamic lexical identifier using a **two-phase lexical index learning** to effectively learn relevance signals.
- Phase 1: Pre-train the model using **keyword-based IDs** to learn the semantics of the corpus.
- Phase 2: Learn **ranking-based document IDs** to reflect query-document relationships.

For inference, GLEN utilizes **collision-free inference** to efficiently rank documents.
- In offline, the **indexing model** generates document identifiers from each document, and similar documents can be mapped to the same identifier.
- In online, the **retrieval model** generates document identifiers from a query, and **collided documents** are efficiently ranked using **identifier weights**.



Experimental Results

- GLEN achieves **state-of-the-art or competitive performance compared to baselines** on benchmark datasets (**NQ320k**).
- GLEN outperforms the **best generative retrieval methods** in the large-scale corpus (**MS MARCO dev**) and zero-shot evaluation setting (**BEIR**).

Type	Model	Natural Questions 320K		
		R@1	R@10	MRR@100
Traditional retrieval	BM25	29.7	60.3	40.2
	DocT5Query	38.0	69.3	48.9
	DPR	50.2	77.7	59.9
	GTR-base	56.0	84.4	66.2
Generative retrieval	DSI	55.2	67.4	59.6
	NCI	66.4	85.7	73.6
	GENRET	68.1	88.8	75.9
	GLEN (Ours)	69.1	86.0	75.4

Model	MS MARCO Dev (MRR@10)	BEIR (nDCG@10)	
		Arguana	NFCorpus
BM25	18.4	31.5	32.5
DocT5Query	27.2	34.9	32.8
DSI	3.1	1.8	11.1
NCI	11.8	0.9	4.3
GENRET	17.4	12.1	12.1
GLEN (Ours)	20.1	17.6	15.9

*Please refer to the paper for more detailed results.

How would you represent G0 in the cell cycle of a neuron?

#	Rel.	Document title	GLEN ID	NCI [1] ID	Keyword ID
1	✓	G0 phase	(#1) phase-phase-cell	(#14) 22-17-10-4	phase-cells-nutri
2		G2 phase	(#2) phase-phase-cell	(#2) 21-28-3-0	phase-phase-cell
3		Cell cycle checkpoint	(#3) point-check-cell	(#9) 1-27-21-1	point-check-cell

Numbers in parentheses: The rank of documents predicted by each model

Check out our paper and code for details!

contact info.
sk1027@skku.edu

Paper Code