BMI: Bottleneck-Minimal Indexing for Generative Document Retrieval

Xin Du* and Kumiko Tanaka-Ishii
Waseda University
https://ml-waseda.jp

Lixin Xiu*
The University of Tokyo

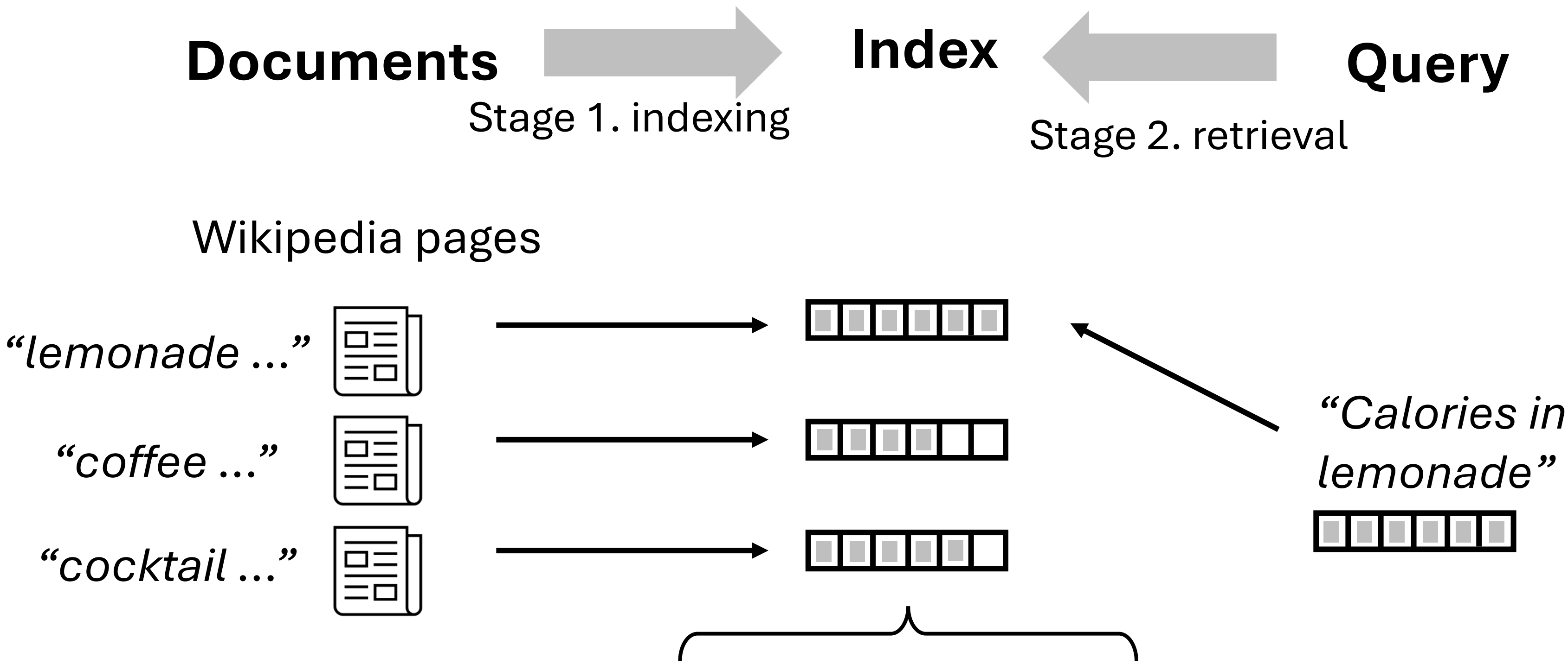
Paper

GitHub

Our lab

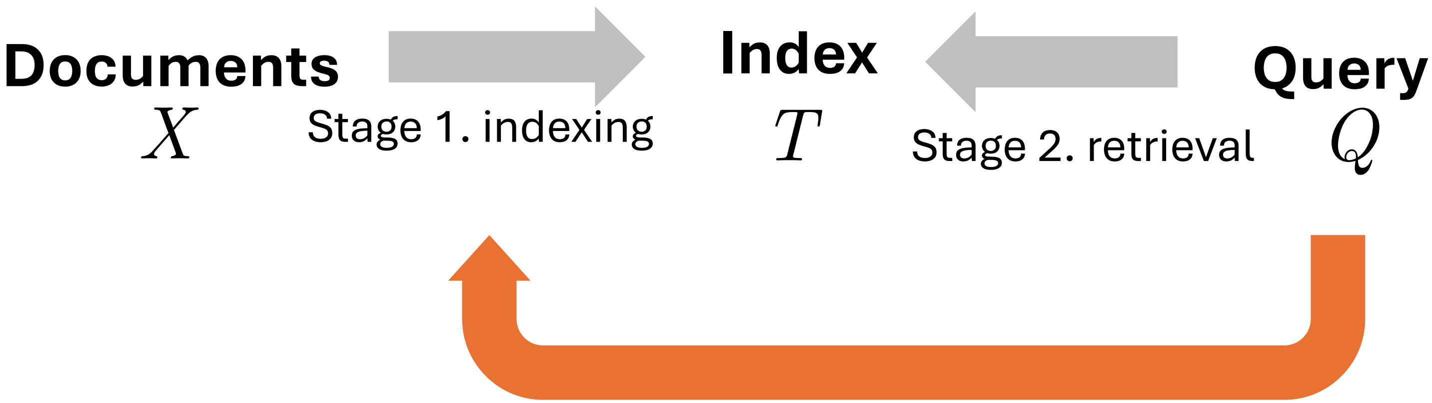
* Equal contribution
Corresponding author

Document Retrieval



- Boolean vectors
- TF-IDF, BM25, ...
- Semantic hashing (2009)
- BERT embedding (2018)

Indexing for Document Retrieval



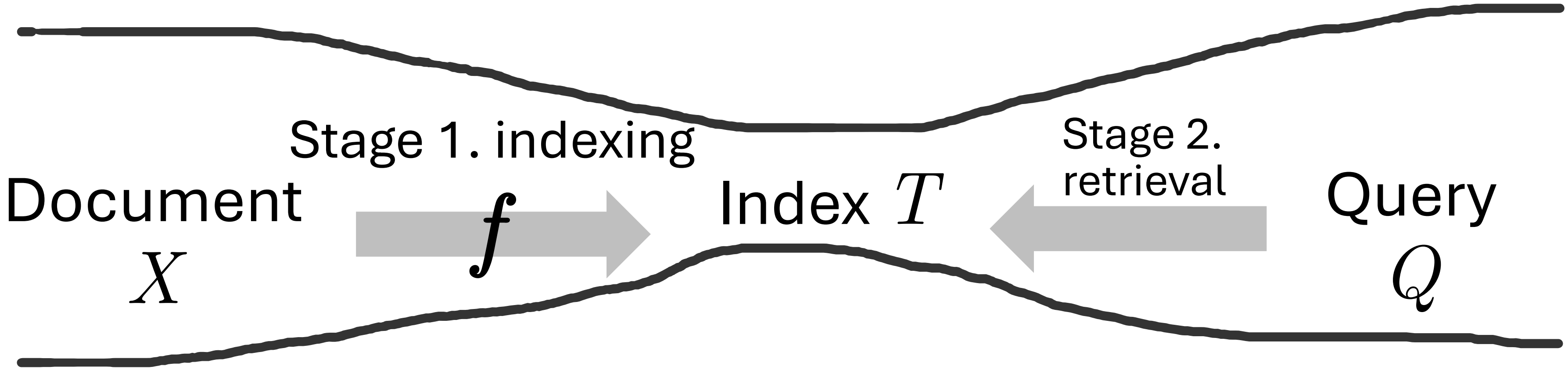
A long-existing idea: **Queries Q should be incorporated** in indexing

But how?

It lacks a theoretical guide.

This Work: An Information Bottleneck Model [1] for Generative Document Retrieval

(GDR explained in the next page)



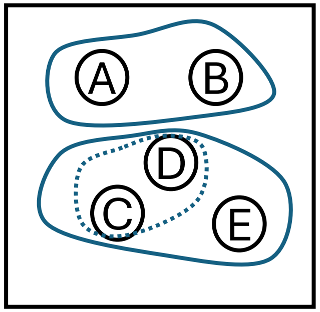
- Retrieval as "data transmission" from query Q to document X.
- Index T as the *"bottleneck"*.

Optimal indexing f^*: $X \mapsto T$ should be ***bottleneck-minimal*.**

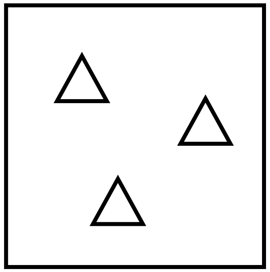
[1] Tishby, Pereira, and Bialek. The information bottleneck method (2000)

Generative Document Retrieval (GDR) [2,3]

document X

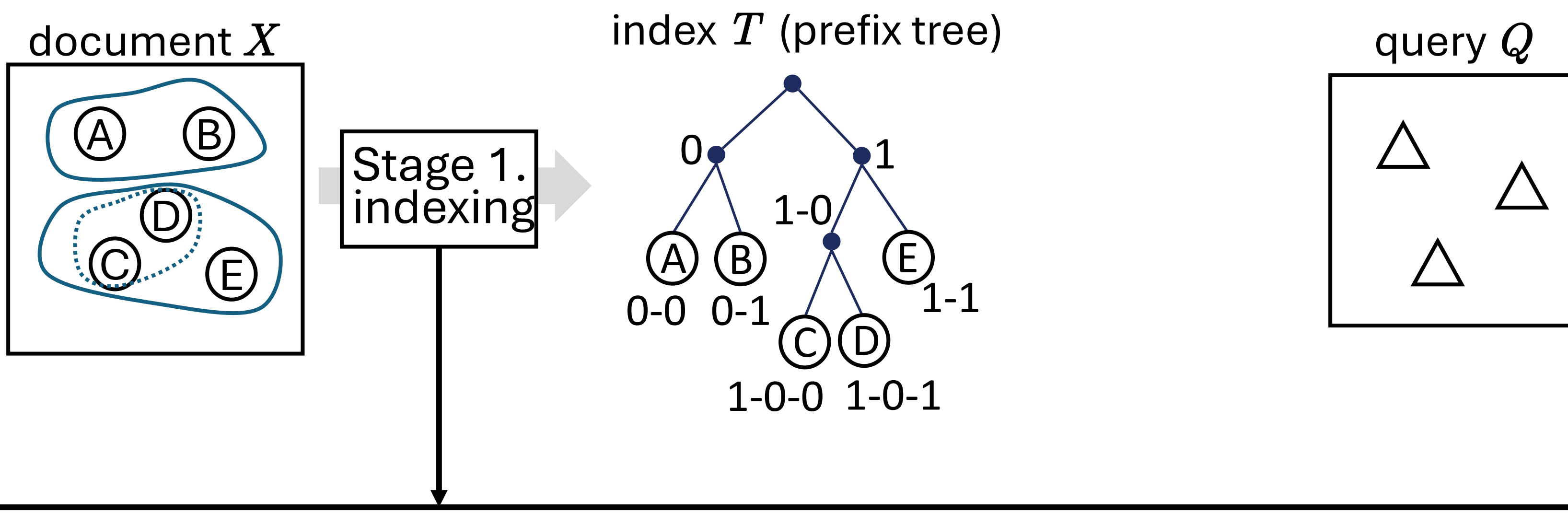


query Q



[2] Cao et al. Auto-regressive entity retrieval. ICLR 2021.
[3] Tay et al. Transformer memory as a differentiable search index. NeurIPS 2022.

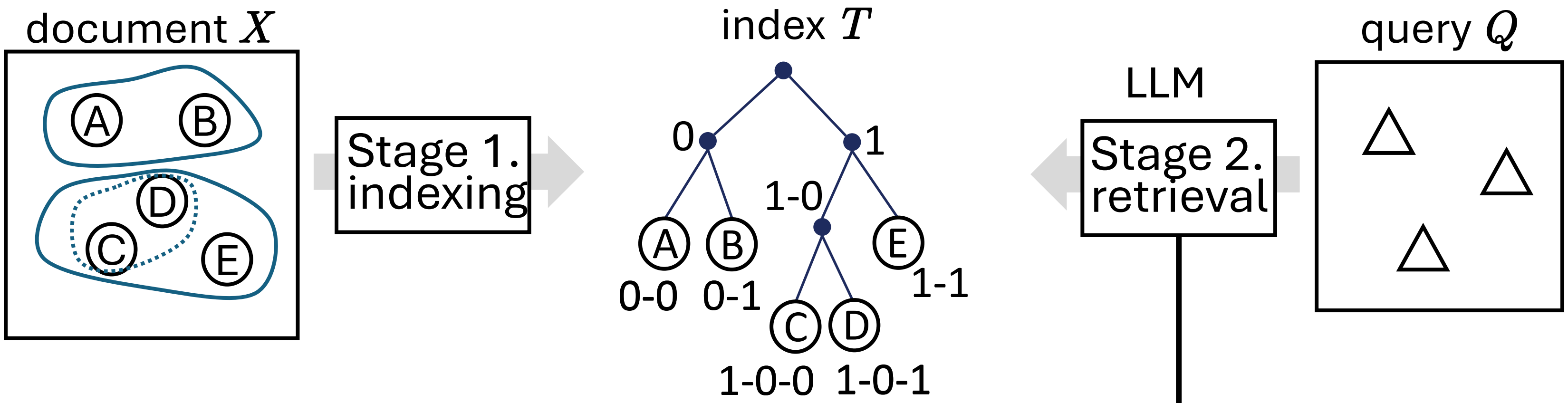
Generative Document Retrieval (GDR)

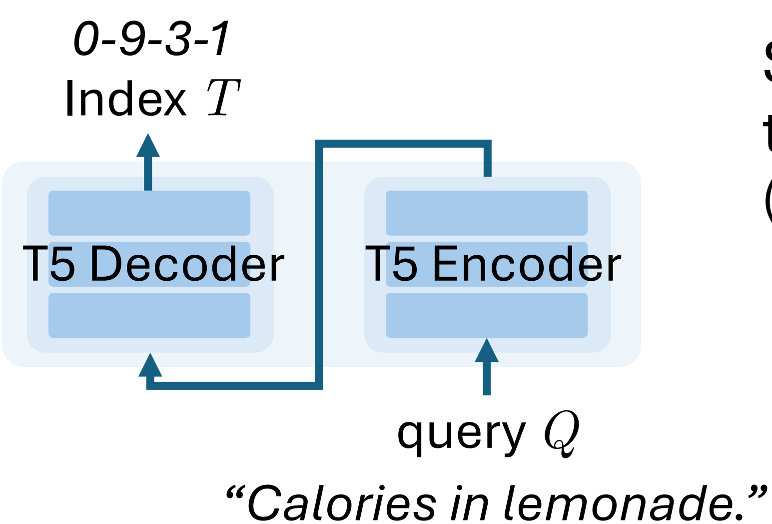


Stage 1.
Hierarchical clustering applied to document vectors. [2-3]

- Good Intuition but lacks theoretical support.

Generative Document Retrieval (GDR)





Stage 2. Train a language model to “translate” a query into index.
(i.e., generate the index digit by digit)

Promising, but limited by the model’s *finite* and often *insufficient* size.

Information Bottleneck (IB) Model for GDR

Document Index Query

$$X \longleftrightarrow T \longleftrightarrow Q$$

$\mathbf{I}(X;T)$ $\mathbf{I}(T;Q)$

IB studies the *tradeoff* by the Lagrangian:

$$L[p(T|X),\beta] = \mathbf{I}(X;T) - \beta\,\mathbf{I}(T;Q) \quad \beta \geq 0$$

assuming Markov chain $T \leftrightarrow X \leftrightarrow Q$

Why is this *tradeoff* essential for GDR ?

Why is this *tradeoff* essential for GDR ?

Document

X

Index

T

Query

Q

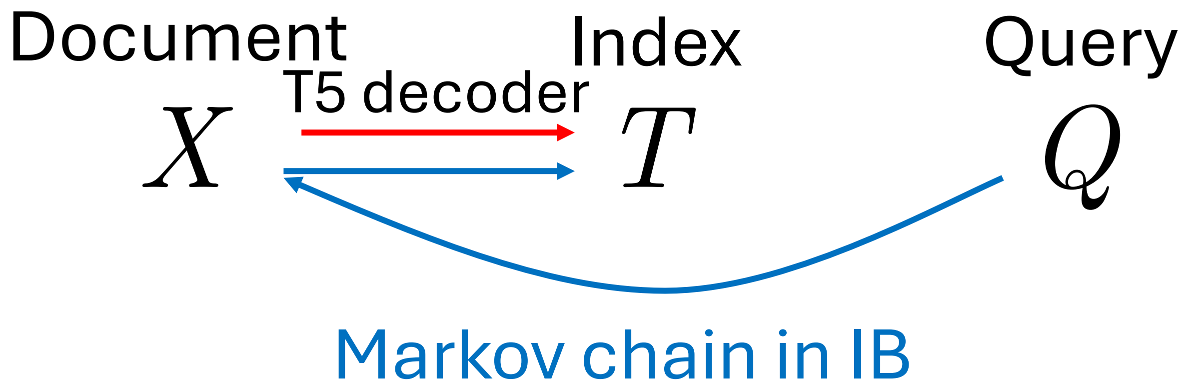
$T \leftarrow Q$: Retrieve by generation

Performance

$\mathbf{I}(T; Q)$

Maximal retrieval accuracy

Why is this *tradeoff* essential for GDR ?



Cost of Memory

$$\mathbf{I}(X;T)$$

How many bits the language model must *"memorize"* per document

Performance

$$\mathbf{I}(T;Q)$$

Maximal retrieval accuracy

Tradeoff: Model Size v.s. Retrieval Performance

minimize

$$L[p(T|X), \beta] = \mathbf{I}(X; T) - \beta\, \mathbf{I}(T; Q)$$

Model size **Performance**

Tradeoff: Model Size v.s. Retrieval Performance

minimize

$$L[p(T|X), \beta] = \mathbf{I}(X; T) - \beta\, \mathbf{I}(T; Q)$$

Model size (T5) **Performance**

Optimal indexing achieves:

highest accuracy for a specific model size, or

Tradeoff: Model Size v.s. Retrieval Performance

minimize

$$L[p(T|X), \beta] = \mathbf{I}(X;\, T) \; - \; \beta\, \mathbf{I}(T;\, Q)$$

Model size (T5) **Performance**

Optimal indexing achieves:

smallest model size for a given accuracy

Tradeoff: Model Size v.s. Retrieval Performance

minimize

$$L[p(T|X), \beta] = \mathbf{I}(X; T) - \beta\, \mathbf{I}(T; Q)$$

Model size (T5) **Performance**

Such optimal indexing under a **limited model-size budgets** is called the ***bottleneck-minimal indexing*** (next page),

NOT the case when a model can be infinitely-large (**unlimited model size**), e.g., a maximum inner-product search model.

This Work: Bottleneck-Minimal Indexing

Our proposed definition of BMI:

An indexing function $f\colon X \mapsto T$ is called an BMI if it maximizes the likelihood function $p(\mathrm{dataset} \mid f)$

$$f^* = \operatorname*{argmax}_f \prod_{\mathrm{doc}\ x} p^*(X{=}x \mid T{=}f(x))$$

Optimal Solution to IB

constant

$$= \operatorname*{argmax}_f \prod_{\mathrm{doc}\ x} \frac{p^*(T{=}f(x) \mid X{=}x)}{p^*(f(x))} p(x)$$

$$= \operatorname*{argmin}_f \sum_{\mathrm{doc}\ x} \mathrm{KL}[p(Q \mid x) \parallel p(Q \mid f(x))]$$

How we acquire f^*

Assume $p(Q|x)$ and $p(Q|f(x))$ to be Gaussian

center vector of Gaussian

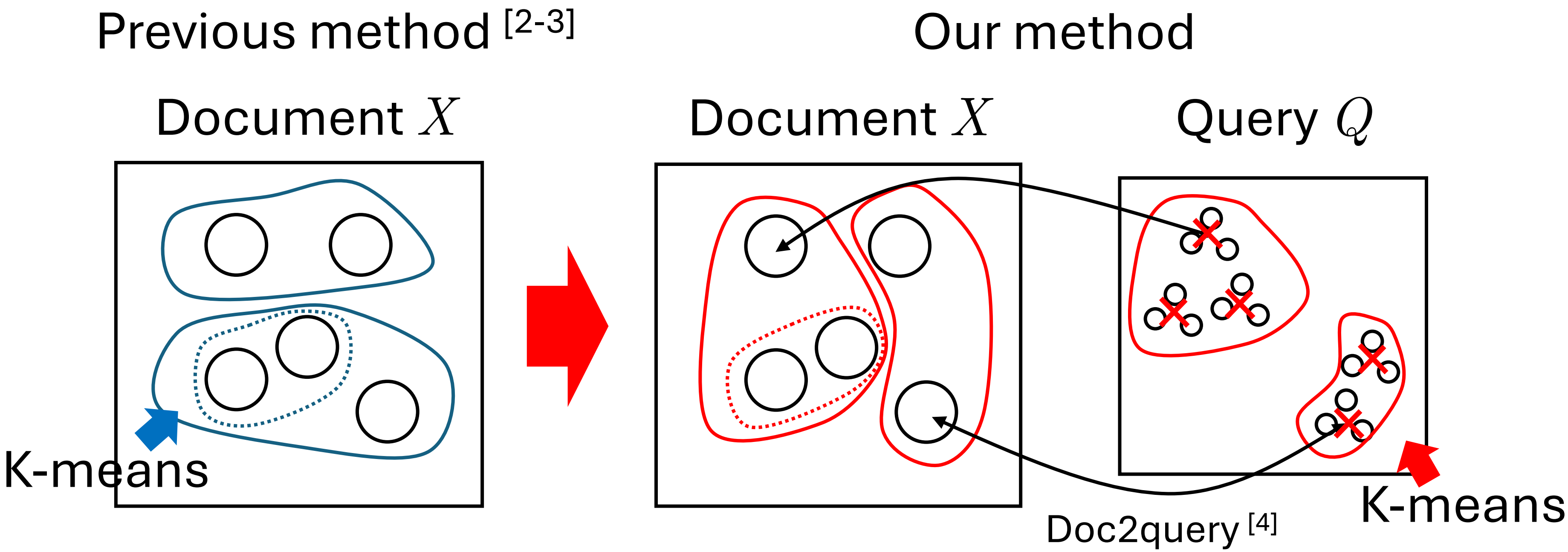
$$f^* = \underset{f}{\operatorname{argmin}} \sum_{\text{doc } x} \| \mathrm{E}[p(Q|x)] - \mathrm{E}[Q|f(x)] \|^2$$

Estimated with a doc2query[4] model

Essentially, we applying k-means clustering to query center vectors $\{\mathrm{E}[p(Q|x)]\}$ instead of document vectors $\{x\}$.

[4] Nogueira and Jimmy Lin. From doc2query to docTTTTTquery (2019).

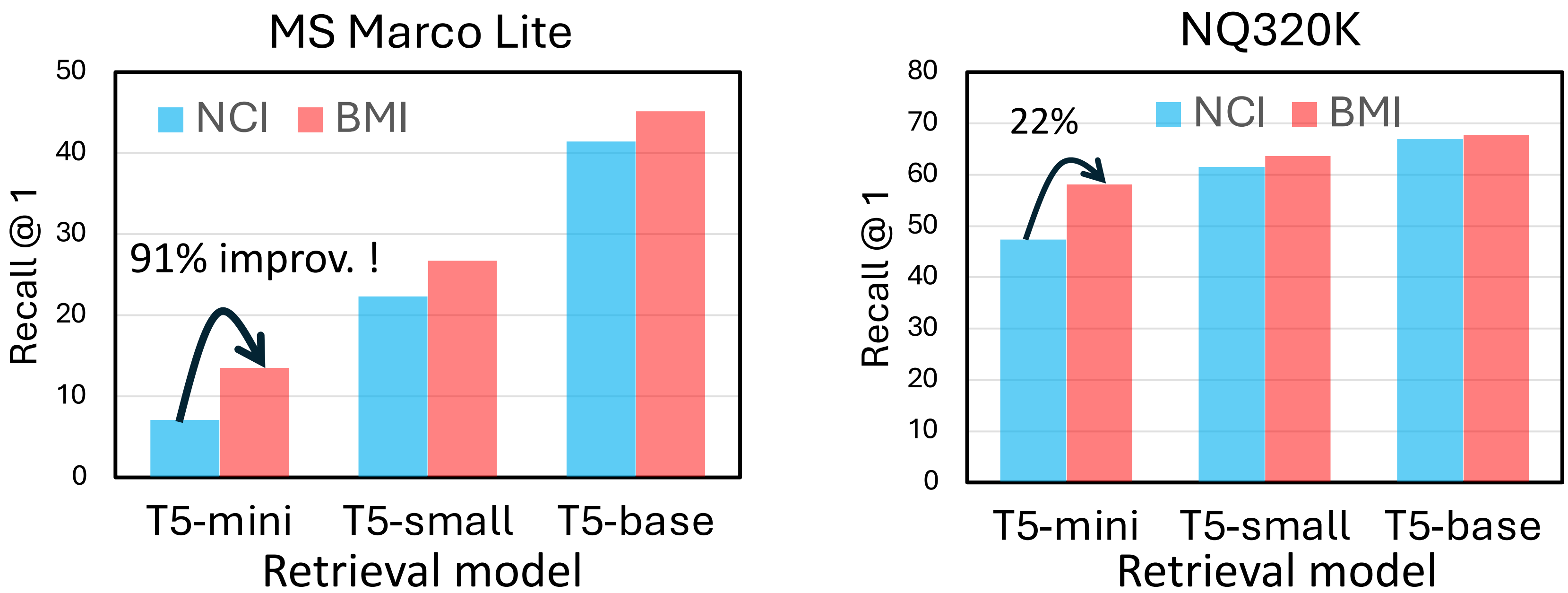
How we acquire f^*



Essentially, we applying k-means clustering to query center vectors $\{\mathrm{E}[p(Q|x)]\}$ instead of document vectors $\{x\}$.

[2] Cao et al. Auto-regressive entity retrieval. ICLR 2021.
[3] Tay et al. Transformer memory as a differentiable search index. NeurIPS 2022.
[4] Nogueira and Jimmy Lin. From doc2query to docTTTTTquery (2019).

Retrieval Accuracy



BMI organize indexing "knowledge" in a much more efficient way than previous GDR methods [2-3] when model size is insufficient, which is common in real-world applications.

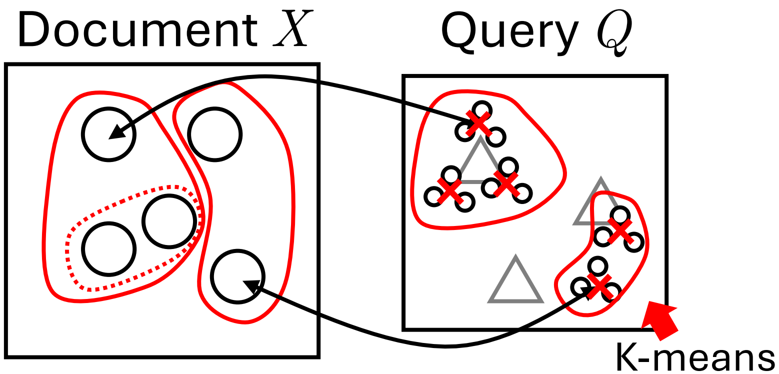
Takeaway

Bottleneck-minimal indexing: an optimal indexing principle under a limited model-size budget.

$$L[p(T|X), \beta] = \mathbf{I}(X; T) - \beta\, \mathbf{I}(T; Q)$$

Model size **Performance**

In GDR, better to watch queries rather than documents.



Information-bottleneck theory worked well to explain how knowledge like indexing can be organized efficiently.

Thank you for listening!

Paper

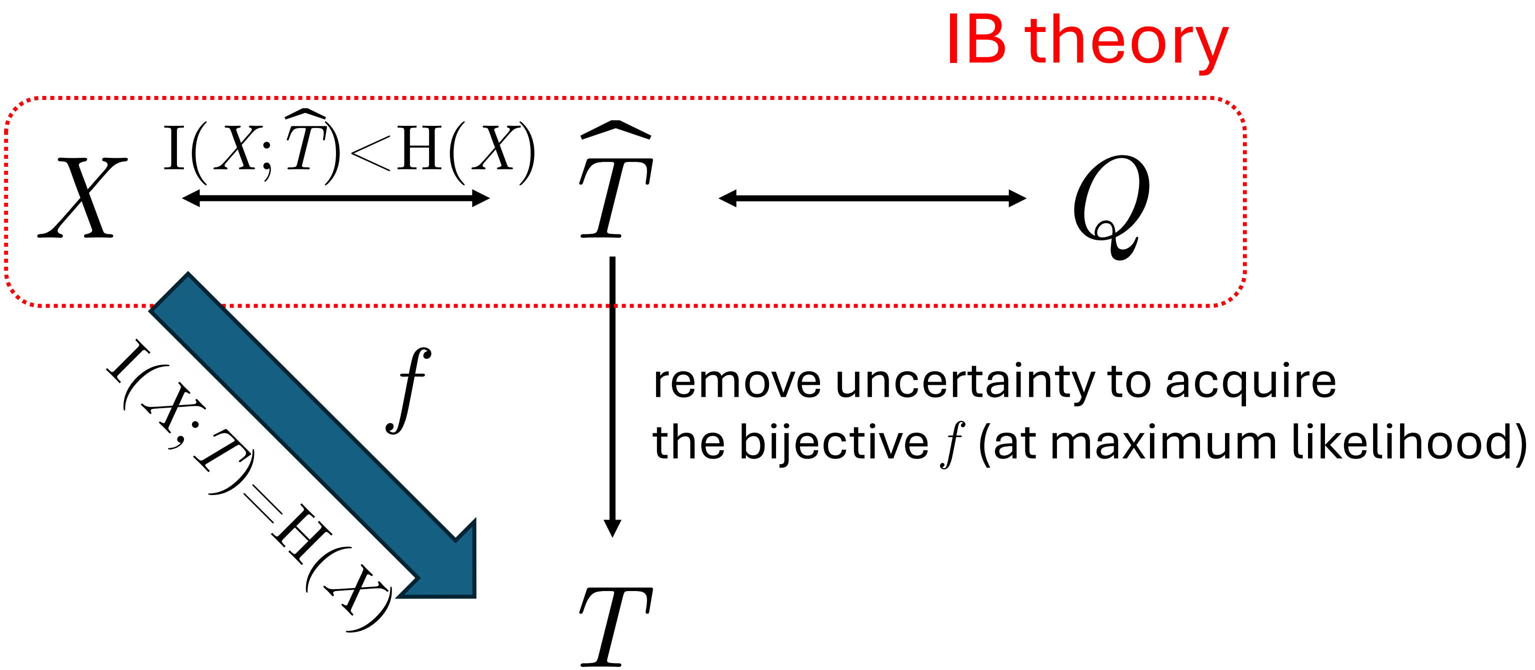
GitHub

Our lab

As f is bijective, isn't $\mathrm{I}(X;T)$ constant ?

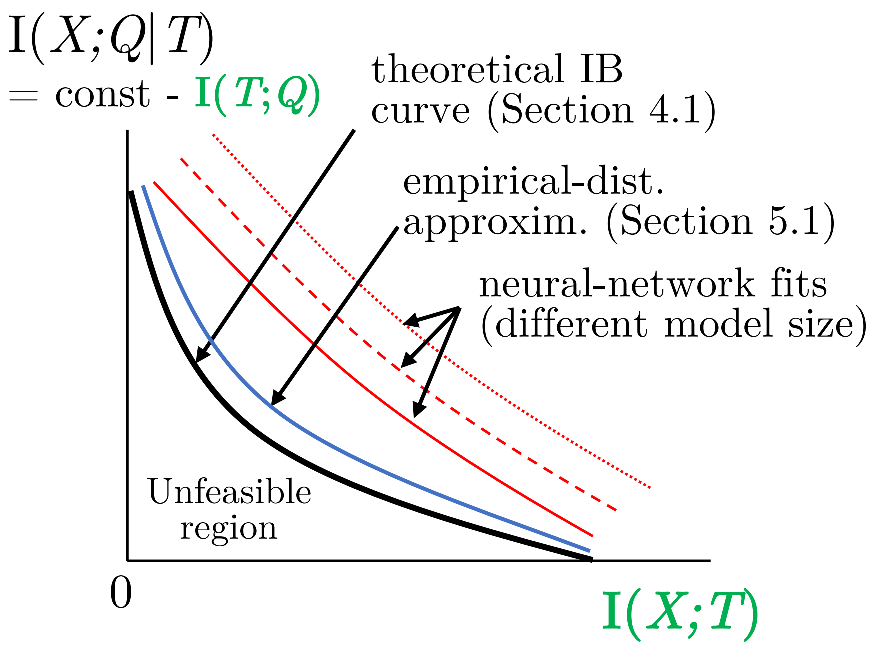
We reused T for two different variables T and $\widehat{T}$ for simpler presentation.

The full scheme is:



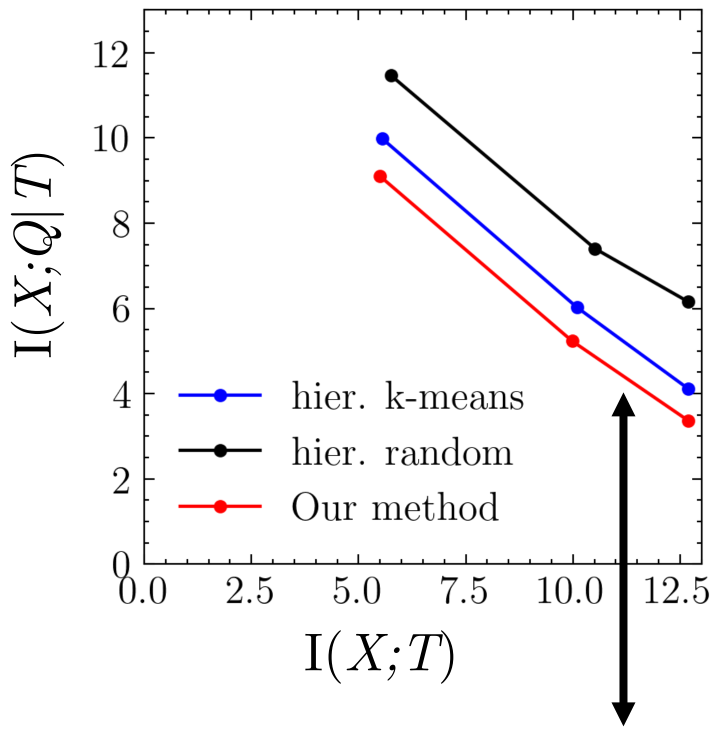
Theoretical & Empirical IB Curves

Theoretical IB curve



IB theory: balance between $\mathrm{I}(X;T)$ and $\mathrm{I}(T;Q)$

Empirical IB curve



Our method is closer to the theoretical IB curve.

Datasets & Settings

	NQ320K	MS Marco Lite
# Documents	109,739	138,457
# Queries (train)	307,373	183,947
# Queries (test	7,830	2,792

Used docT5query (base) [5] for query generation.

Retrieval model training: used the implementation of NCI [3]

The change from NCI [3] is only the indexing $f: X \mapsto T$