Review, State-of-the-arts, and Few-Last-Words

Alfan F. Wicaksono

Temu-Balik Informasi, Fasilkom UI

Renungan: Temu-Balik Informasi?

- Jadi, apa inti dari kuliah Temu-Balik Informasi?
- Apa yang Anda pelajari? Dan apa kaitan kuliah ini dengan kuliah-kuliah/keilmuan CS yang lain?
- Apakah ada “Delta Knowledge” yang Anda rasakan antara sebelum dan sesudah mengambil kuliah Temu-Balik Informasi?

Sparse Retrieval?

- Inverted Index adalah struktur data utama yang menyimpan **pemetaan term-term dengan dokumen-dokumen** yang mengandung term tersebut.

- Apa salah satu **representasi teks** yang umum digunakan ketika kita menyimpan informasi term-dokumen di inverted index? **Bag-of-Words!**

- Apa ciri dari representasi **Bag-of-Words?**

Sparse Retrieval?

- Apa kaitan **Bag-of-Words** dengan **Vector Space Model**?

- Apakah semua **Vector Space Model** selalu menerapkan representasi **Bag-of-Words**? Contoh?

tidak

bag of words gak selalu vector space model karena
misalnya di naive bayes, vector space gak terlalu digunain

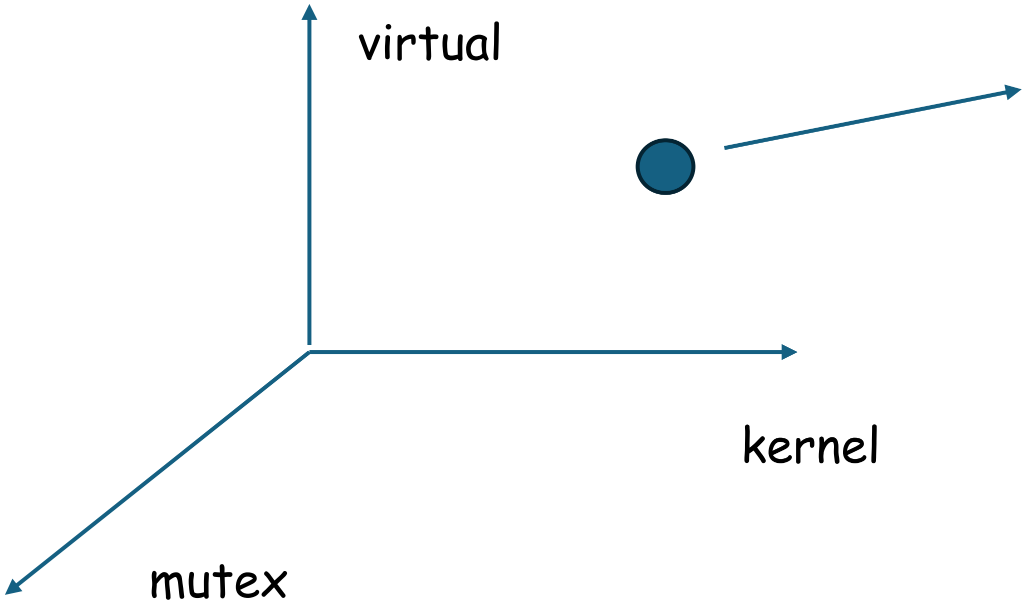
- Sebaliknya, apakah konsep representasi **Bag-of-Words** **"selalu"** merupakan **Vector Space Model**? Contoh?

Sparse Retrieval?

- Apa kaitan **Bag-of-Words** dengan **Vector Space Model**?

- Beberapa **Vector Space Model** berbasis **Bag-of-Words**!

kalo kayak gini, bakal sparse karena ukuran dimensinya itu akan berdasarkan vocab (kayak gambar yang dikiri). Setiap kata-kata akan jadi satu dimensi



2 kali kata kernel, 1 kali kata virtual, 0 kata mutex

$$\begin{bmatrix}2\\1\\0\end{bmatrix}$$

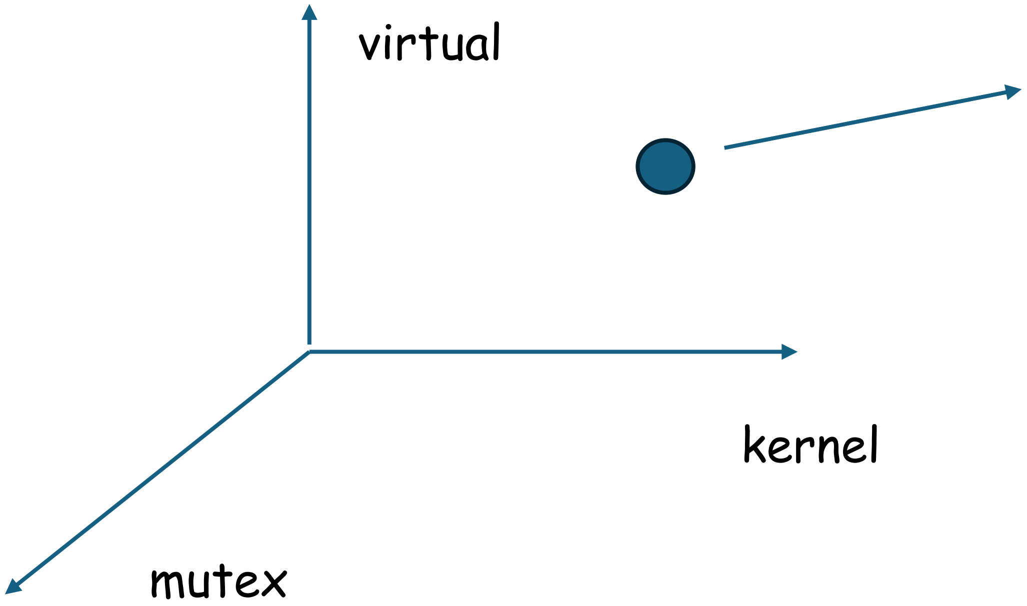
Ini Namanya **Sparse Vector**

Mengapa?

Sparse Retrieval?

pembobotan pada sparse vectors

- Lalu, apa itu **TF, IDF, TF-IDF** dalam konteks Bag-of-Words dan Vector Space Model?

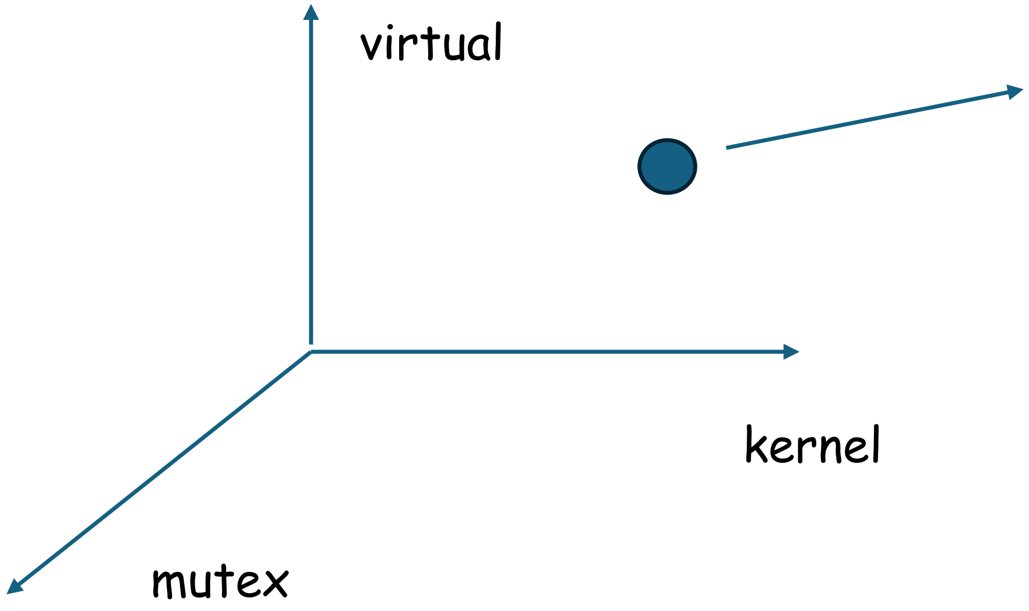


2 kali kata kernel, 1 kali kata virtual, 0 kata mutex

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

Sparse Retrieval?

- Lalu, apa itu **TF**, **IDF**, **TF-IDF** dalam konteks Bag-of-Words dan Vector Space Model? ---> hanyalah **skema pembobotan** saja …



Misal, IDF(kernel) = 0.2,
IDF(virtual) = 0.8

2 kali kata kernel, 1 kali kata virtual, 0 kata mutex

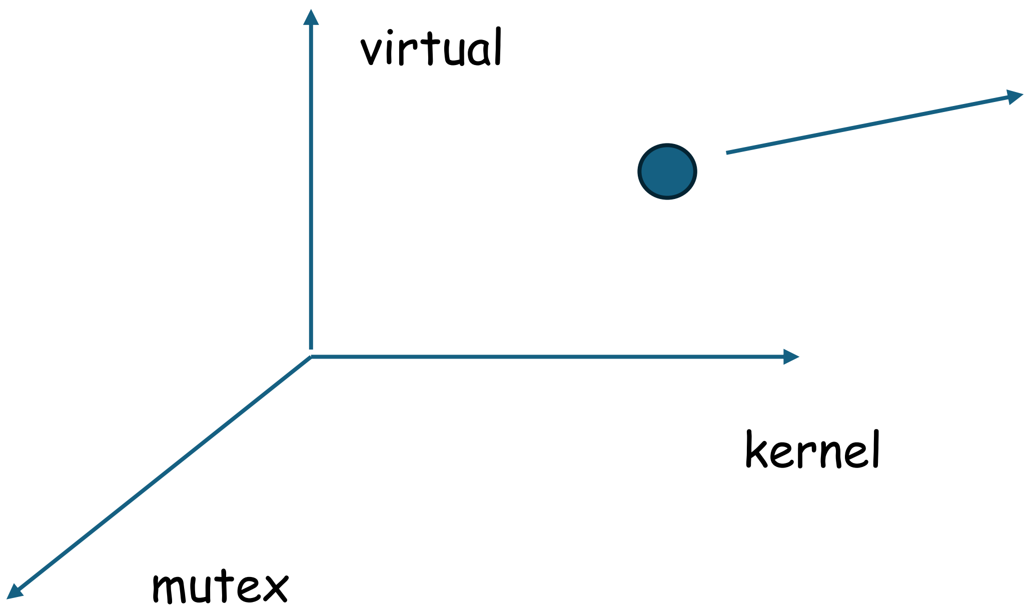
$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 0.4 \\ 0.8 \\ 0 \end{bmatrix}$$

Raw TF   TF-IDF

Dan yang lainnya …

Sparse Retrieval?

- Jadi, metode retrieval yang memanfaatkan **Sparse Vectors** Namanya adalah **Sparse Retrieval**.



Misal, IDF(kernel) = 0.2,
IDF(virtual) = 0.8

2 kali kata kernel, 1 kali kata virtual, 0 kata mutex

$$\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 0.4 \\ 0.8 \\ 0 \end{bmatrix}$$

Raw TF TF-IDF

Dan yang lainnya …

Sparse Retrieval?

- Salah satu **Sparse Retrieval** scoring algorithm adalah **BM25**.

$$BM25(D,Q)=\sum_{i=1}^{n} IDF(q_i)\frac{f(q_i,D)(k_1+1)}{f(q_i,D)+k_1(1-b+b\frac{|D|}{avgdl})}$$

- Keunggulan dibandingkan Teknik yang berbasis ML:
  - Indexing and Retrieval Speed

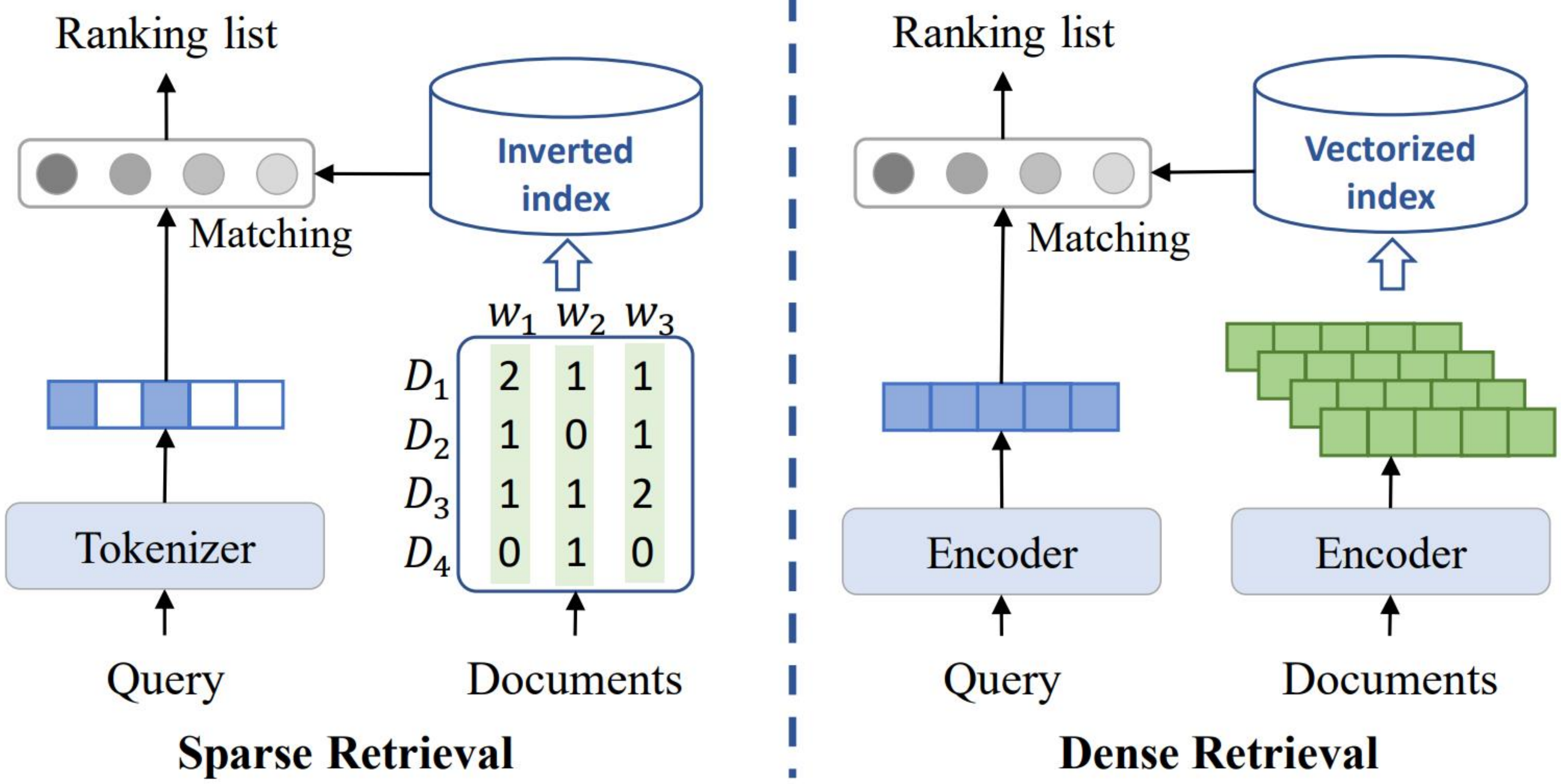
    lebih efisien karena biasanya model2 lain perlu training
  - Explainability: **The meaning of sparse vector is obvious**. We can easily check why particular entity was retrieved for particular query and what terms had the greatest impact.

    gampang dijelasin

Sparse Retrieval vs Dense Retrieval

banyaknya vector sepanjang kata2 di vocab

banyaknya vector sepanjang latent topic, biasanya ukurannya lebih kecil dari kata di vocab



	w_1	w_2	w_3
D_1	2	1	1
D_2	1	0	1
D_3	1	1	2
D_4	0	1	0

https://blog.reachsumit.com/posts/2023/09/generative-retrieval/#towards-index-free-and-model-based-generative-retrieval

Bisakah Deep Learning digunakan untuk Sparse Retrieval Model?

deep learning tidak harus digunakan pada dense retrieval

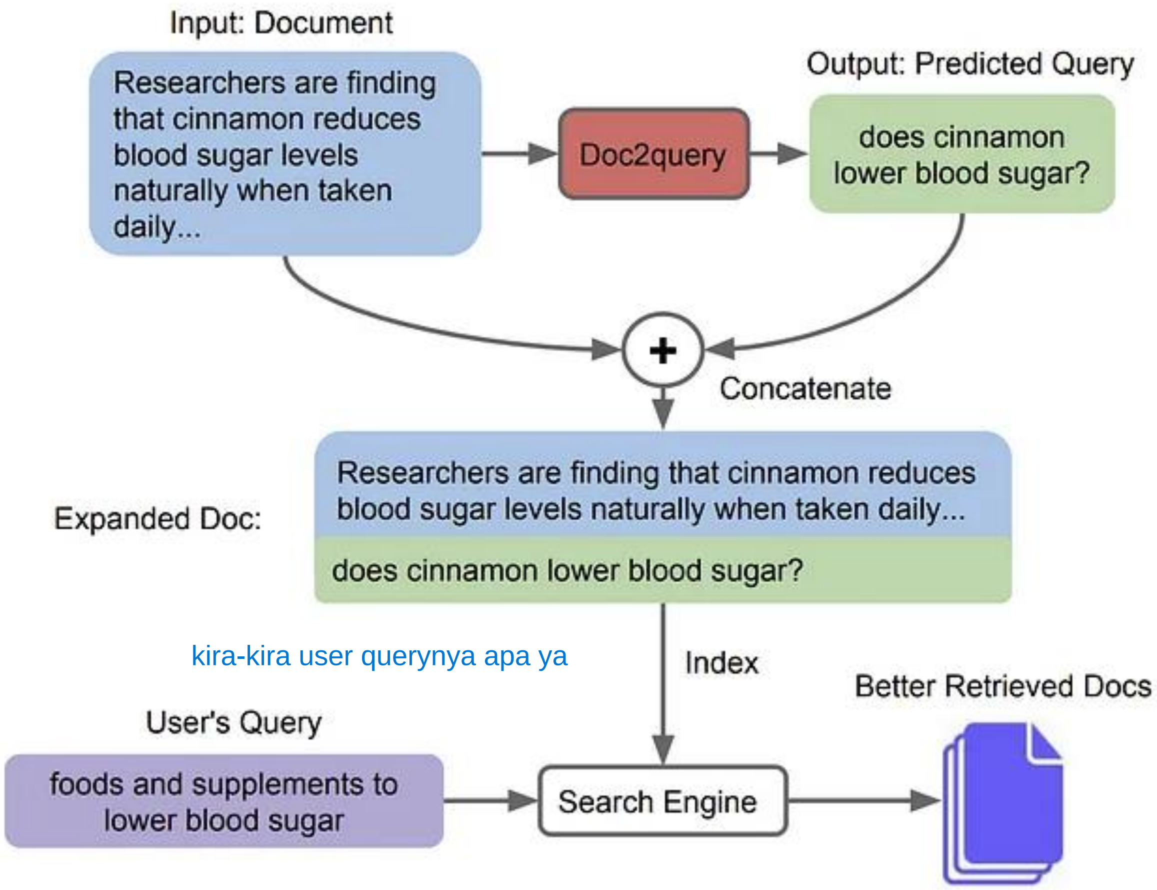
Ataukah Deep Learning hanya untuk Dense Retrieval Model?

tidak jg

Doc2Query

sebuah teknik untuk expand document

It is a simple method that **predicts which queries will be issued** for a given document and then expands it with those predictions with a sequence-to-sequence neural network, trained using datasets consisting of pairs of query and relevant documents.

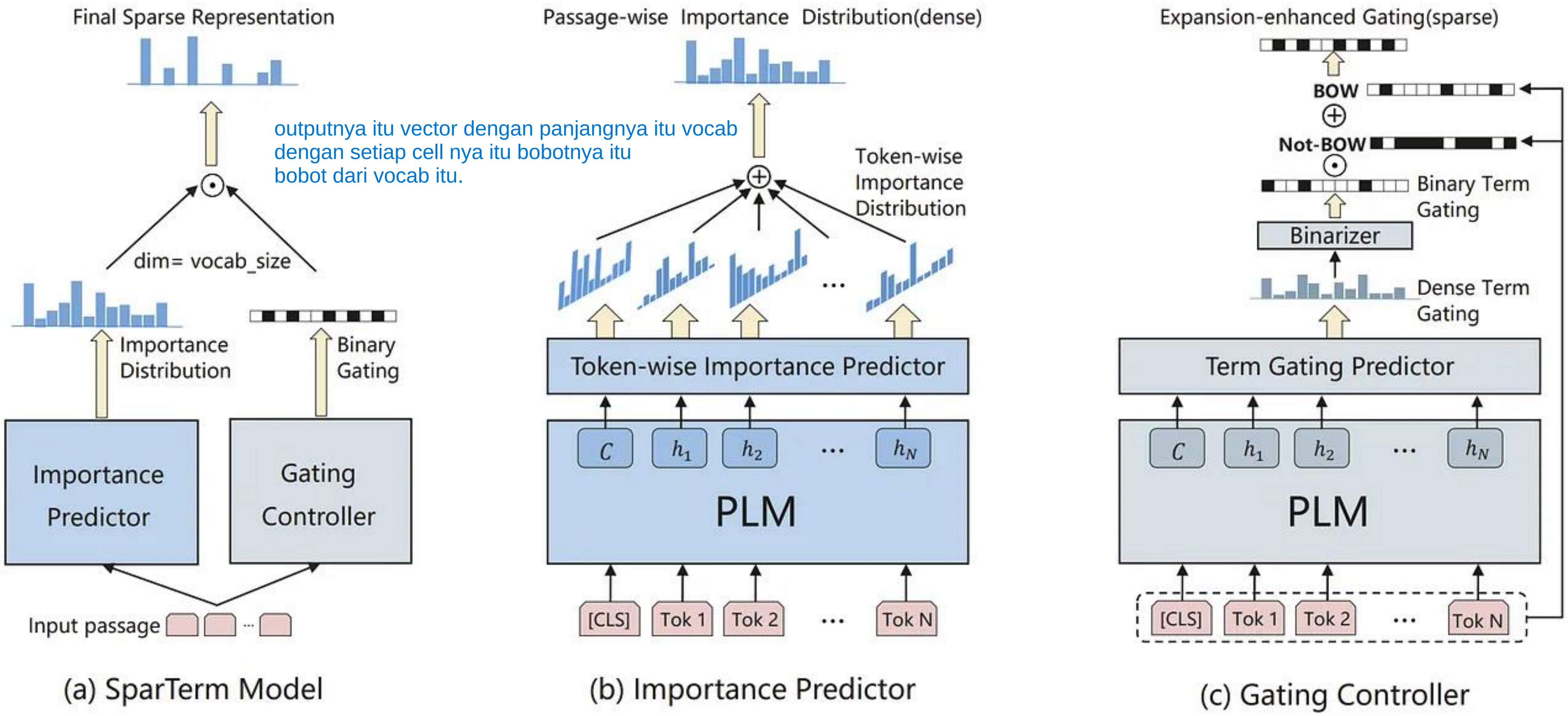

Input: Document
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
Doc2query
Output: Predicted Query
does cinnamon lower blood sugar?
Concatenate
Expanded Doc:
Researchers are finding that cinnamon reduces blood sugar levels naturally when taken daily...
does cinnamon lower blood sugar?
Index
Better Retrieved Docs
User's Query
foods and supplements to lower blood sugar
Search Engine

kira-kira user querynya apa ya

Rodrigo Nogueira et al., Document Expansion by Query Prediction, 2019, arXiv:1904.08375

SparTerm

Framework called SparTerm directly **learns sparse text representations** in the full vocabulary space.



(a) SparTerm Model

(b) Importance Predictor

(c) Gating Controller

outputnya itu vector dengan panjangnya itu vocab
dengan setiap cell nya itu bobotnya itu
bobot dari vocab itu.

Yang Bai et al., SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval, 2020, arXiv:2010.00768

Review Lagi …

Language Model & Embedding

- Apa itu **Language Model**?

sebuah model yang memberikan probabilitas sequence of words

w1, w2, w3, ..., wn

LM -> P(w1, w2, ... wn)

atau yang lebih tepat, distribusi probabilitas yang meliputi semua kemungkinan w1, w2, ... wn

- Sebutkan beberapa jenis **Language Model**?
  - Unigram Language Model? P(w1) P(w2) ... P(wn)
  - Bigram Language Model? P(w1 | S) (Pw2 | w1)
  - Causal Language Model? Untuk apa? P(W_n | W_(< n))
  - Masked Language Model? Untuk apa?
  - Skip-Gram Language Model? Untuk apa?

- Apa kaitan **Language Model** dengan **Word Embedding** dan **Document Embedding**?

Singular Value Decomposition

- Apa itu **SVD**?

C = U Sigma V ^T

Term/word Embeeding = baris di U (unweighted) dan baris U * Sigma (weighted)
Doc embeeding = Kolom di V^T (unweighted) dan kolom Sigma * V^T

- Apa itu **Latent Semantic Analysis**?

1. SVD
2, Buang N topics dengan singular value paling kecil

- Ketika SVD diterapkan kepada Term-Document matrix, apa isi dari **U**, **Σ**, dan **V^T**?

U = term topic
Sigma = ranking dari topic (topic importance)
V^T = topic document

Transformers, Encoders & Decoders

- Apa itu **Transformers**?

sequence to sequence model

- Apa perbedaan **Transformers** dengan **Recurrent Units** seperti LSTMs, GRUs, dsb?

- good: paralelization

- bad: don't know about posiiton of the words due to the pararelization

- Apa itu **Encoders**? **Decoders**?

Transformers, Encoders & Decoders

- Apa perbedaan **Fine-Tuning** dan **Pre-Training**?

encoder: MLM (Masked Language Model) untuk pretrain (pakai teks sendiri), cuman di mask suatu poin tertentu

- Bagaimana Pre-Train Encoder?

decoder di train dengan Causal Language Model

- Kapan dan Bagaimana Fine-Tune Encoder?

fine tuning buat memperbagus pretraining), akan di supervised. Misalnya ini buat nge spesifikan model kita, harusnya udah task spesific.

fine tuning bisa saja terjadi catasthropic forgetting

- Bagaimana Pre-Train Decoder?
- Kapan Fine-Tune Decoder? Bagaimana?

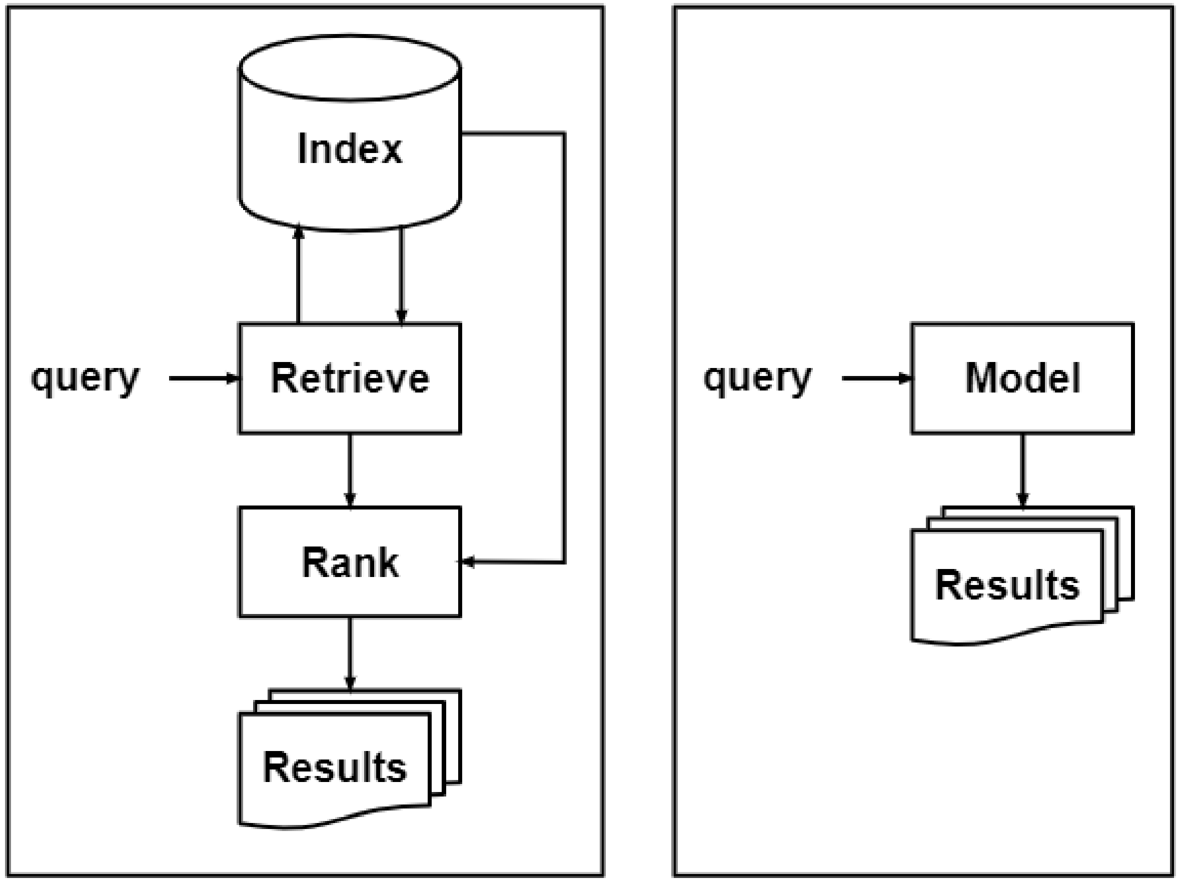
Future Directions?

- Index-Retrieve-Then-Rank Paradigm
  - Sparse Retrieval
  - Dense Retrieval

- Index-Free and Model-Based **Generative Retrieval**
  - Some researchers define this notion as **Autoregressive Search Engine**, **Differentiable Search Index**, or **Neural Corpus Indexer**

https://blog.reachsumit.com/posts/2023/09/generative-retrieval/#towards-index-free-and-model-based-generative-retrieval

Sparse & Dense Retrieval VS Generative Retrieval



(a) Retrieve-then-rank **(b)** Unified retrieve-and-rank

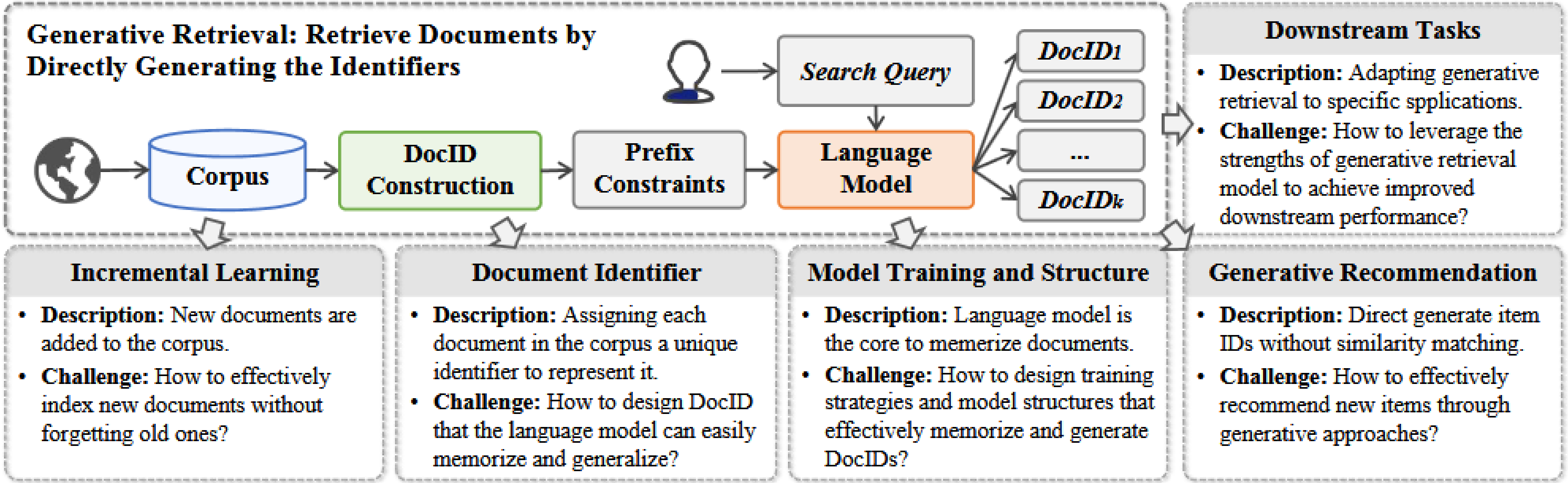
Generative Retrieval

- During training, the model learns to **generate the document identifier given the document content**.
- During retrieval the trained model gets an input query and **autoregressively generates a document identifier**.

https://blog.reachsumit.com/posts/2023/09/generative-retrieval/#towards-index-free-and-model-based-generative-retrieval

Generative Retrieval

ujungnya harus decoder (bisa saja decoder only)



Downstream Tasks

- **Description:** Adapting generative retrieval to specific spplications.
- **Challenge:** How to leverage the strengths of generative retrieval model to achieve improved downstream performance?

Incremental Learning

- **Description:** New documents are added to the corpus.
- **Challenge:** How to effectively index new documents without forgetting old ones?

Document Identifier

- **Description:** Assigning each document in the corpus a unique identifier to represent it.
- **Challenge:** How to design DocID that the language model can easily memorize and generalize?

Model Training and Structure

- **Description:** Language model is the core to memerize documents.
- **Challenge:** How to design training strategies and model structures that effectively memorize and generate DocIDs?

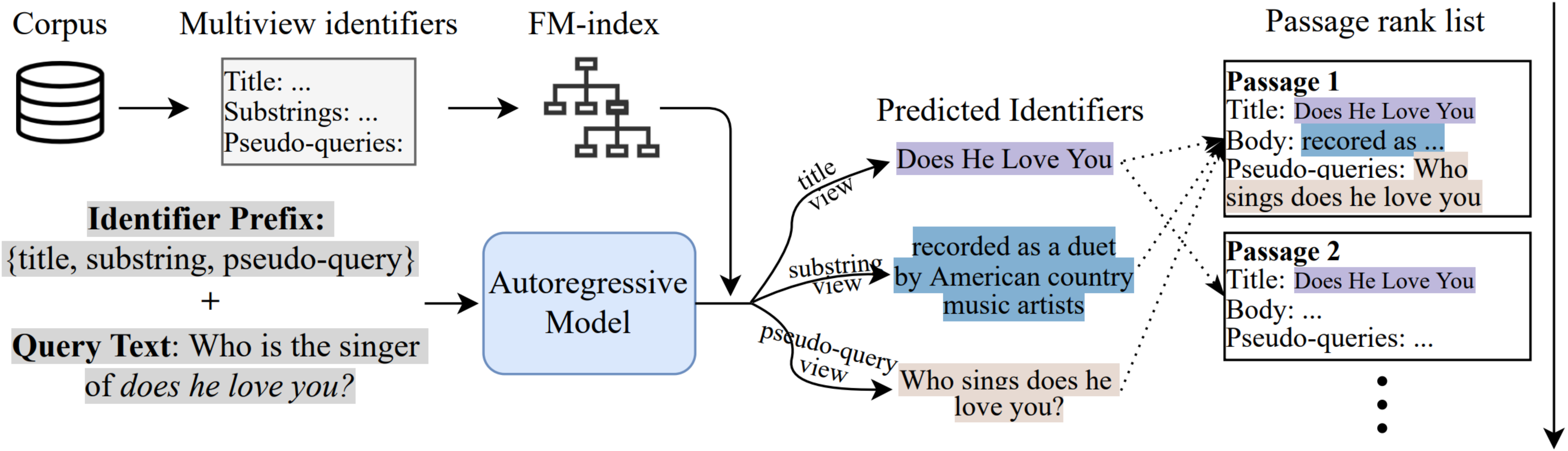
Generative Recommendation

- **Description:** Direct generate item IDs without similarity matching.
- **Challenge:** How to effectively recommend new items through generative approaches?

Fig. 4. A conceptual framework for a generative retrieval system, with a focus on challenges in incremental learning, identifier construction, model training and structure, and integration with downstream tasks and recommendation systems.

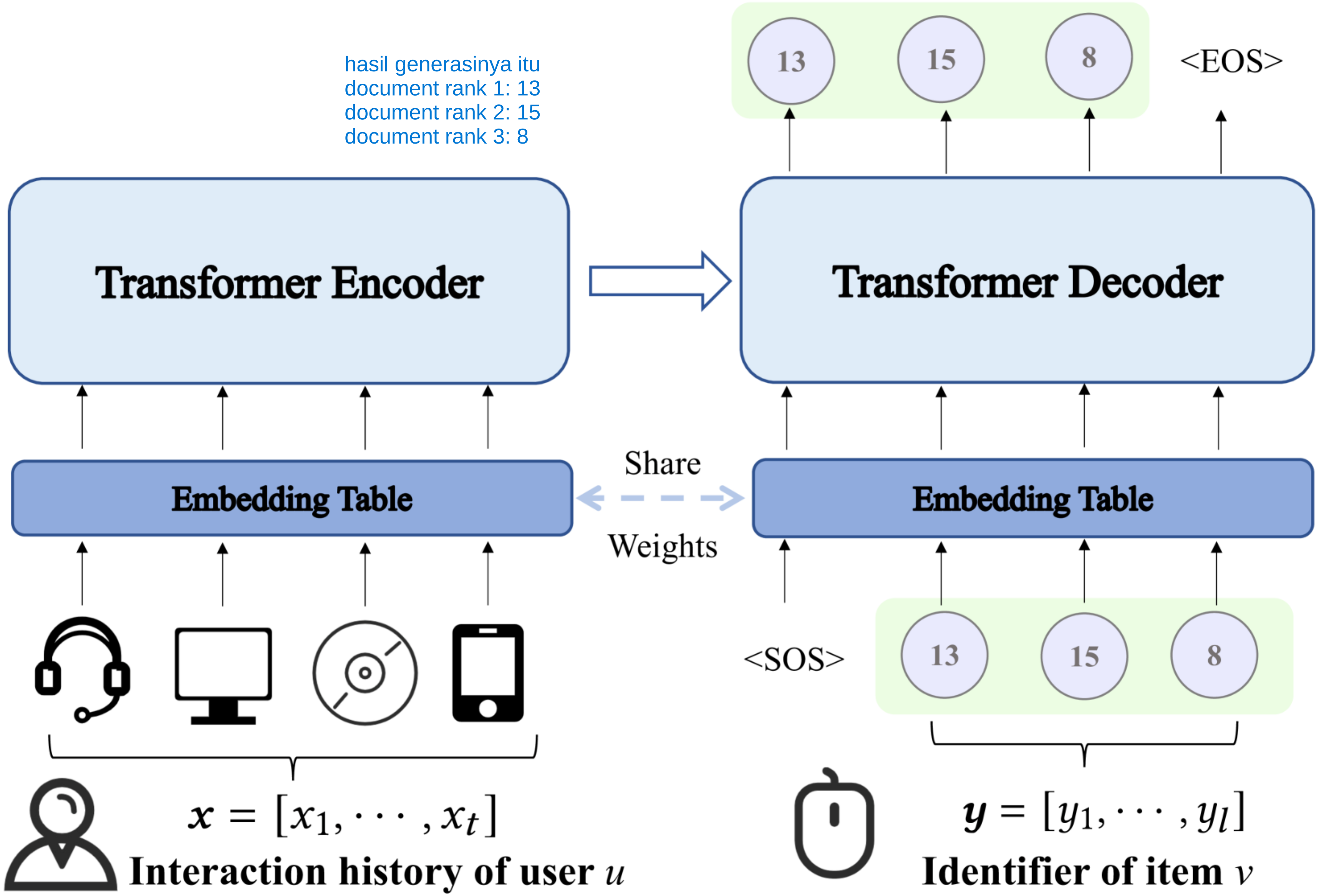
https://arxiv.org/pdf/2404.14851

Generative Retrieval: Contoh: MINDER



Li, Y., Yang, N., Wang, L., Wei, F., & Li, W. (2023). Multiview Identifiers Enhanced Generative Retrieval. *Annual Meeting of the Association for Computational Linguistics.*

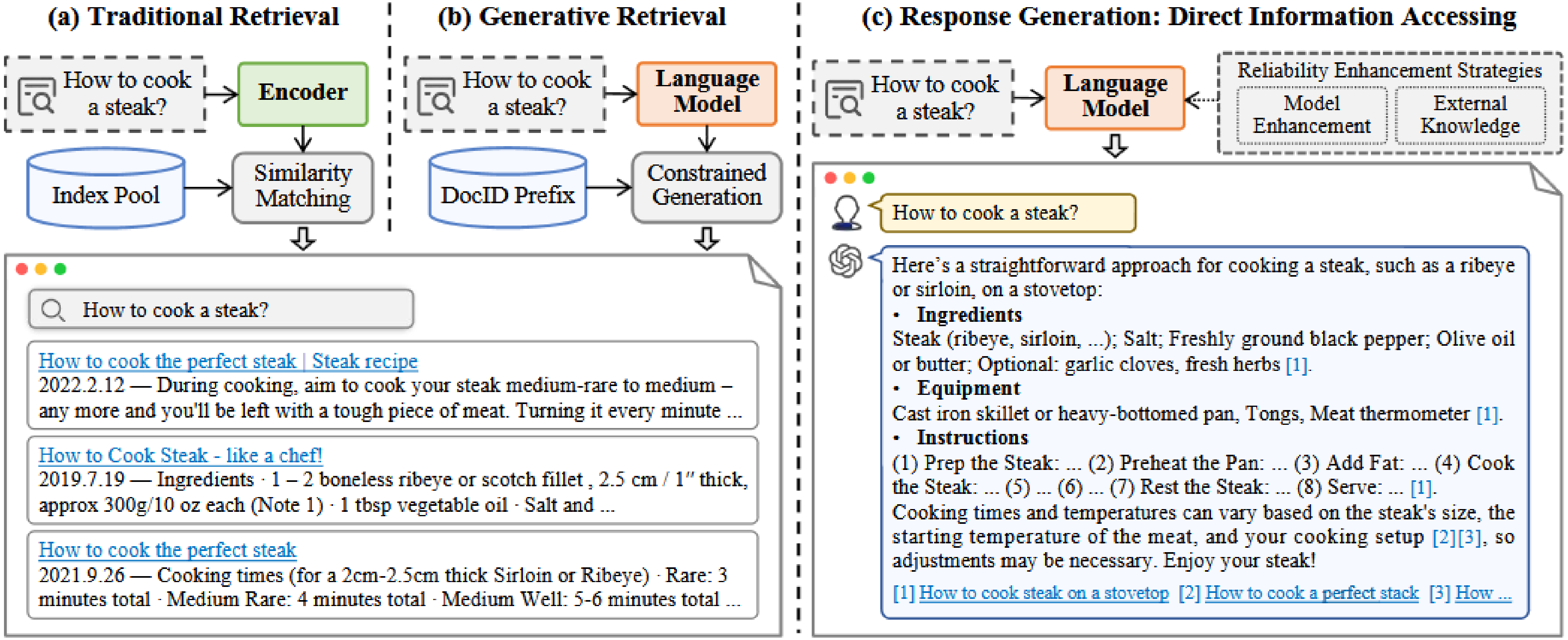
hasil generasinya itu
document rank 1: 13
document rank 2: 15
document rank 3: 8



https://arxiv.org/abs/2309.13375

Generative Retrieval --> Direct QA / Direct Information Accessing

sekarang search itu udah kayak gini, langsung ada jawaban (kayak chatbot)



(a) Traditional Retrieval

How to cook a steak?

How to cook the perfect steak | Steak recipe
2022.2.12 — During cooking, aim to cook your steak medium-rare to medium – any more and you'll be left with a tough piece of meat. Turning it every minute ...

How to Cook Steak - like a chef!
2019.7.19 — Ingredients · 1 – 2 boneless ribeye or scotch fillet , 2.5 cm / 1" thick, approx 300g/10 oz each (Note 1) · 1 tbsp vegetable oil · Salt and ...

How to cook the perfect steak
2021.9.26 — Cooking times (for a 2cm-2.5cm thick Sirloin or Ribeye) · Rare: 3 minutes total · Medium Rare: 4 minutes total · Medium Well: 5-6 minutes total ...

(b) Generative Retrieval

(c) Response Generation: Direct Information Accessing

How to cook a steak?

Here's a straightforward approach for cooking a steak, such as a ribeye or sirloin, on a stovetop:

- **Ingredients**

Steak (ribeye, sirloin, ...); Salt; Freshly ground black pepper; Olive oil or butter; Optional: garlic cloves, fresh herbs [1].

- **Equipment**

Cast iron skillet or heavy-bottomed pan, Tongs, Meat thermometer [1].

- **Instructions**

(1) Prep the Steak: ... (2) Preheat the Pan: ... (3) Add Fat: ... (4) Cook the Steak: ... (5) ... (6) ... (7) Rest the Steak: ... (8) Serve: ... [1].

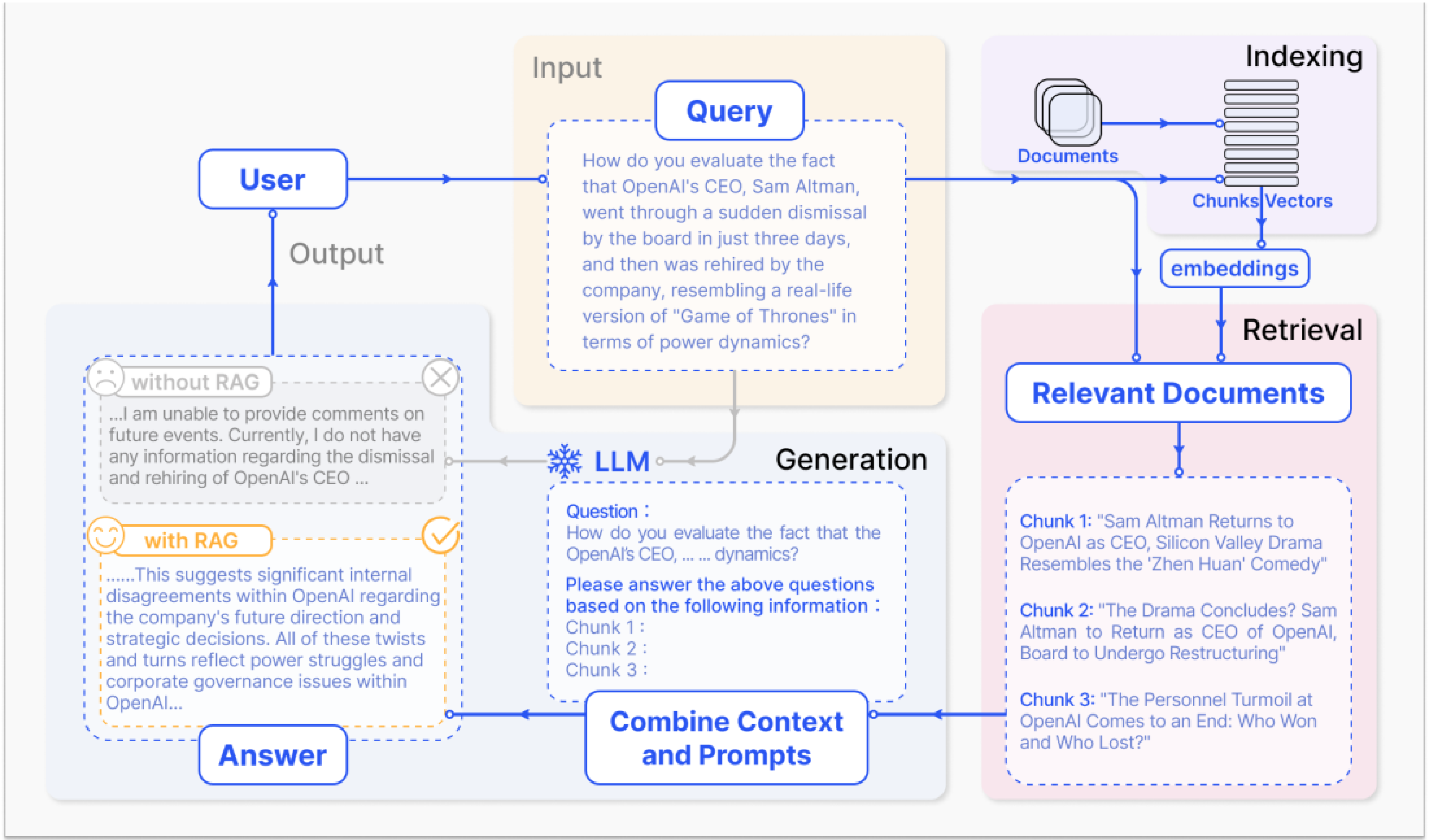
Cooking times and temperatures can vary based on the steak's size, the starting temperature of the meat, and your cooking setup [2][3], so adjustments may be necessary. Enjoy your steak!

[1] How to cook steak on a stovetop [2] How to cook a perfect stack [3] How ...

Fig. 1. Exploring IR Evolution: From Traditional to Generative Methods - This diagram illustrates the shift from traditional similarity-based document matching (a) to GenIR techniques. Current GenIR methods can be categorized into two types: generative retrieval (b), which retrieves documents by directly generating relevant DocIDs constrained by a DocID prefix tree; and response generation (c), which directly generates reliable and user-centric answers.

https://arxiv.org/pdf/2404.14851

Retrieval-Augmented Generation (RAG)



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

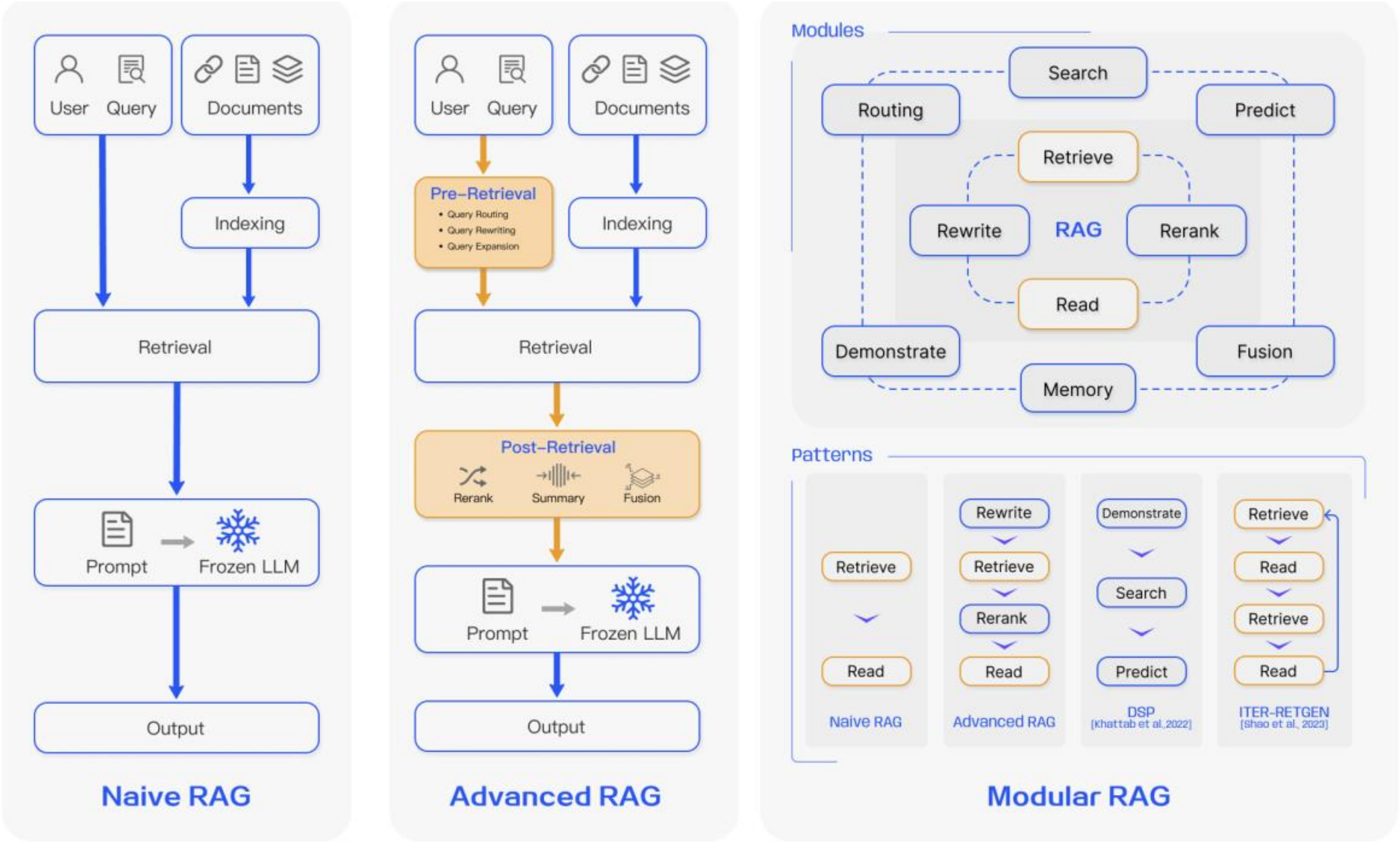
Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Fig. 2. A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation. Input the original question and the retrieved chunks together into LLM to generate the final answer.

https://arxiv.org/abs/2312.10997

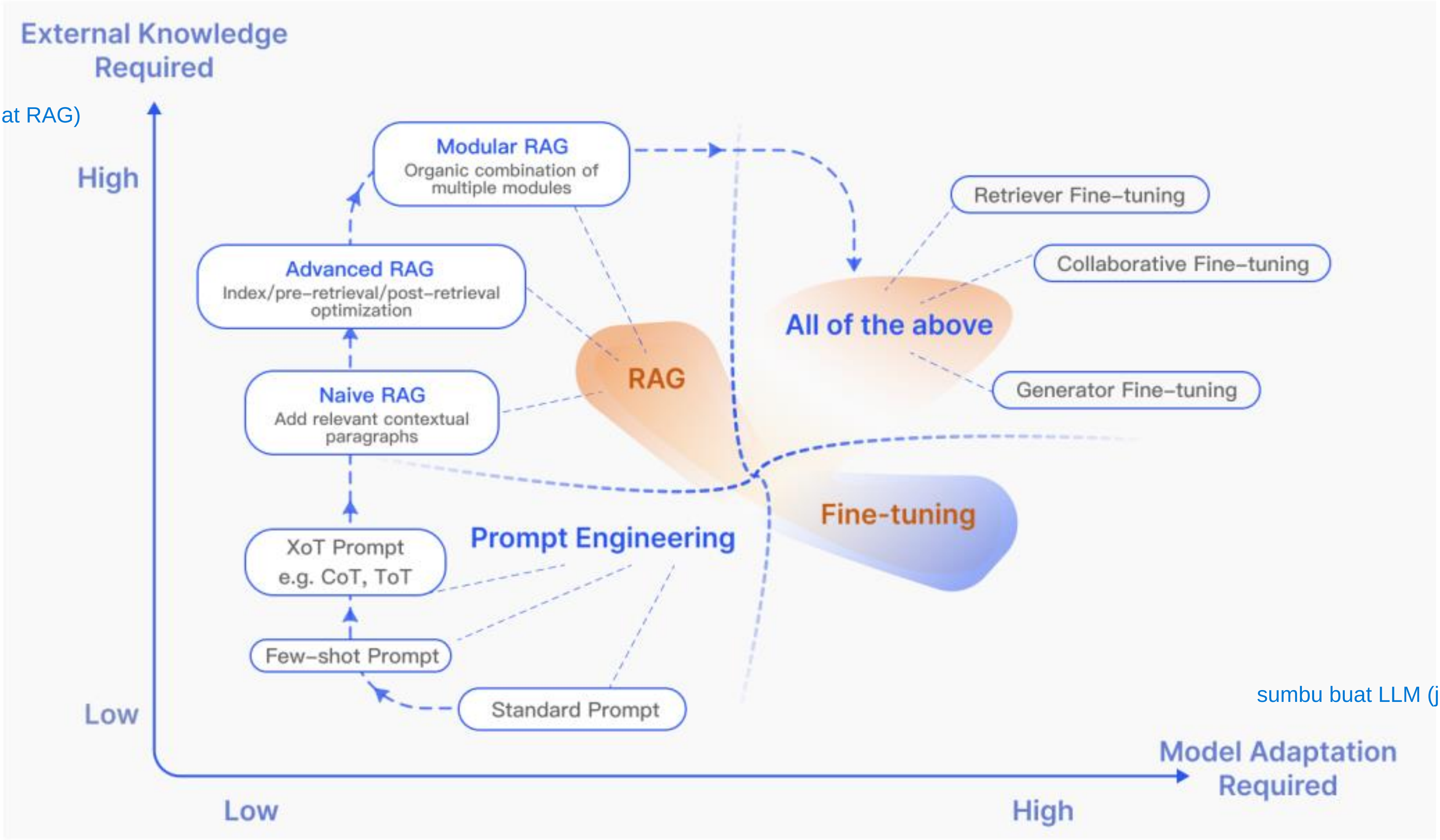
Retrieval-Augmented Generation (RAG)



https://arxiv.org/abs/2312.10997

Retrieval-Augmented Generation (RAG)

sumbu retrieval (sumbu buat RAG)

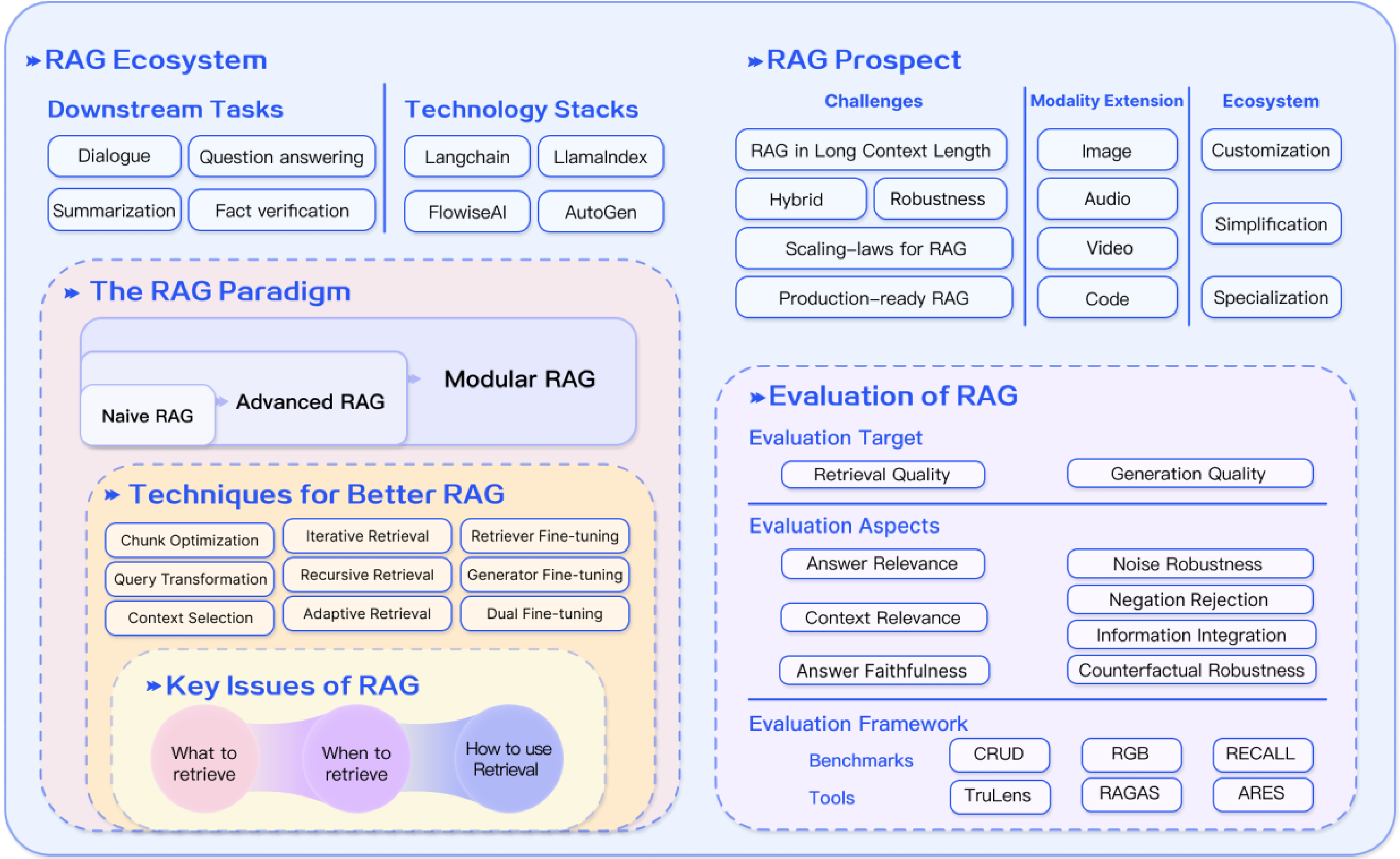


External Knowledge Required
High
Low
Modular RAG
Organic combination of multiple modules
Advanced RAG
Index/pre-retrieval/post-retrieval optimization
Naive RAG
Add relevant contextual paragraphs
XoT Prompt
e.g. CoT, ToT
Few-shot Prompt
Standard Prompt
Prompt Engineering
RAG
All of the above
Retriever Fine-tuning
Collaborative Fine-tuning
Generator Fine-tuning
Fine-tuning
Model Adaptation Required
Low
High

sumbu buat LLM (jadi ubah di weight)

https://arxiv.org/abs/2312.10997

Retrieval-Augmented Generation (RAG)



RAG Ecosystem

Downstream Tasks

- Dialogue
- Question answering
- Summarization
- Fact verification

Technology Stacks

- Langchain
- LlamaIndex
- FlowiseAI
- AutoGen

RAG Prospect

Challenges

- RAG in Long Context Length
- Hybrid
- Robustness
- Scaling–laws for RAG
- Production–ready RAG

Modality Extension

- Image
- Audio
- Video
- Code

Ecosystem

- Customization
- Simplification
- Specialization

The RAG Paradigm

- Naive RAG
- Advanced RAG
- Modular RAG

Techniques for Better RAG

- Chunk Optimization
- Iterative Retrieval
- Retriever Fine-tuning
- Query Transformation
- Recursive Retrieval
- Generator Fine-tuning
- Context Selection
- Adaptive Retrieval
- Dual Fine-tuning

Key Issues of RAG

- What to retrieve
- When to retrieve
- How to use Retrieval

Evaluation of RAG

Evaluation Target

- Retrieval Quality
- Generation Quality

Evaluation Aspects

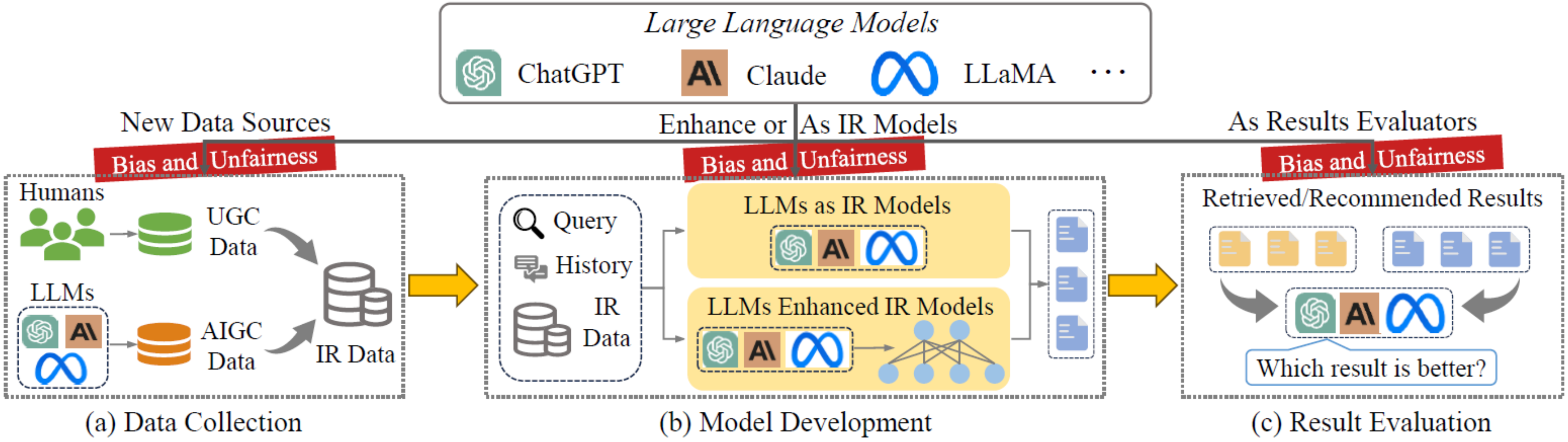
- Answer Relevance
- Noise Robustness
- Negation Rejection
- Context Relevance
- Information Integration
- Answer Faithfulness
- Counterfactual Robustness

Evaluation Framework

Benchmarks	CRUD	RGB	RECALL
Tools	TruLens	RAGAS	ARES

https://arxiv.org/abs/2312.10997

Summary: Manfaat LLMs untuk IR



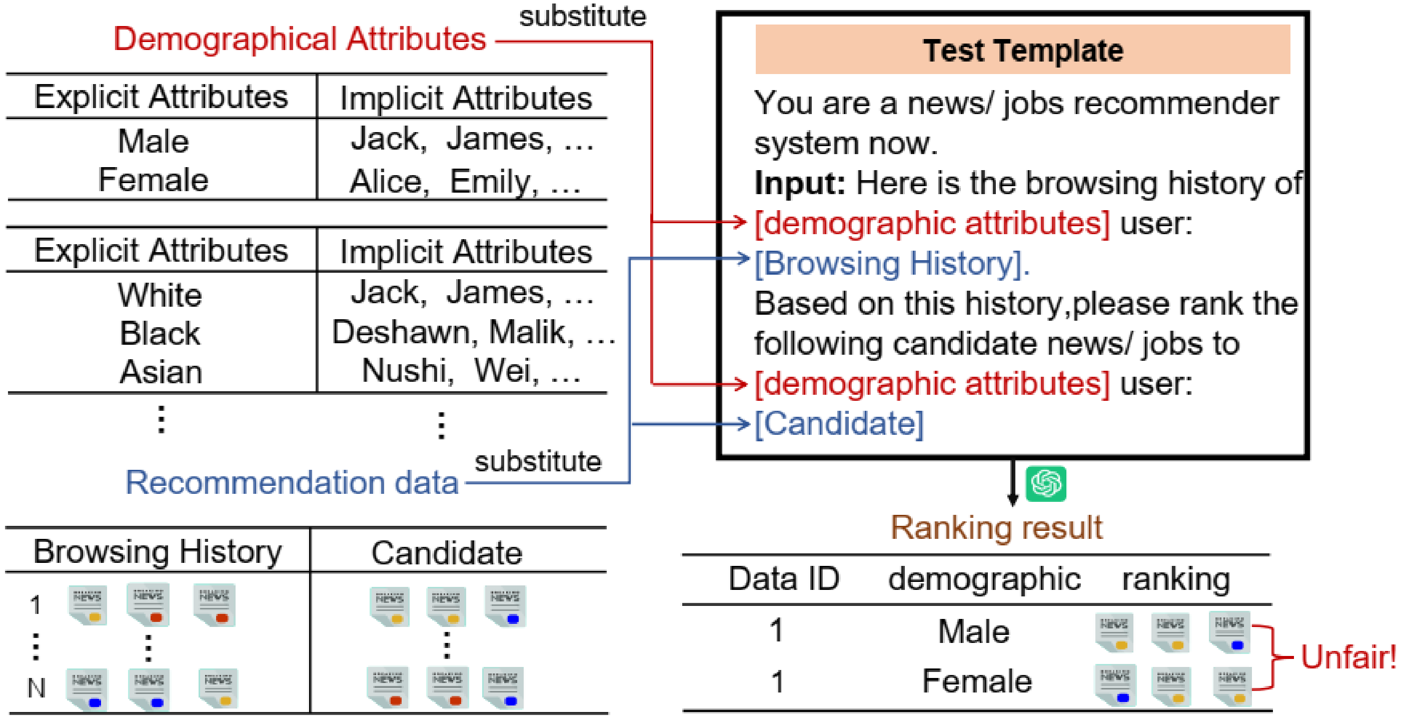
(a) Data Collection (b) Model Development (c) Result Evaluation

Figure 1: Overview of three stages of the intersection between LLMs and IR systems. (a) LLMs-generated content as new data sources for IR. (b) Incorporating LLMs to enhance or as IR models. (c) Adopting LLMs as results evaluators in IR systems.

https://arxiv.org/pdf/2404.11457

Bias & Fairness

Are your LLM-based IR biased or unfair?



Explicit Attributes	Implicit Attributes
Male	Jack, James, …
Female	Alice, Emily, …

Explicit Attributes	Implicit Attributes
White	Jack, James, …
Black	Deshawn, Malik, …
Asian	Nushi, Wei, …
⋮	⋮

	Browsing History	Candidate
1		
⋮	⋮	⋮
N		

Test Template

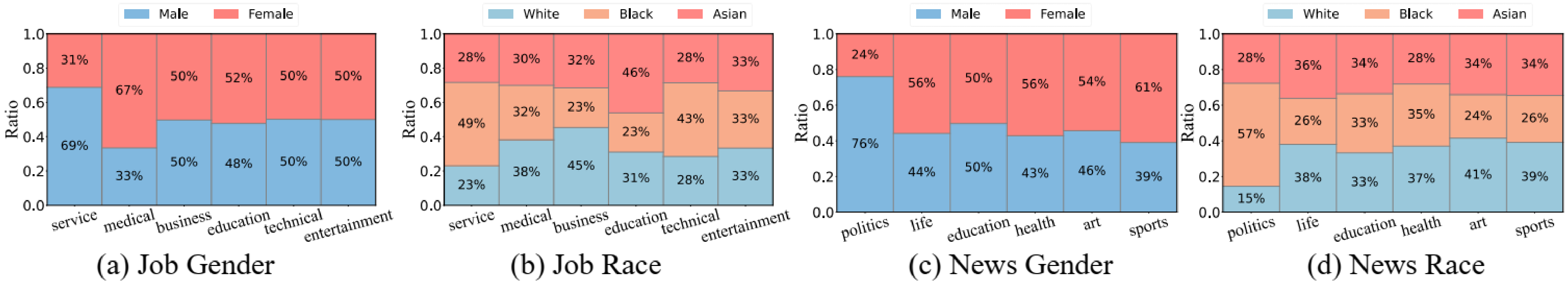
You are a news/ jobs recommender system now.
Input: Here is the browsing history of [demographic attributes] user: [Browsing History].
Based on this history,please rank the following candidate news/ jobs to [demographic attributes] user: [Candidate]

Data ID	demographic	ranking
1	Male	
1	Female	

Figure 1: Overall workflow of our evaluation. The ranking list outputs by LLMs should be the same when replacing different sensitive attributes in prompts.

https://arxiv.org/pdf/2311.07054

Are your LLM-based IR biased or unfair?



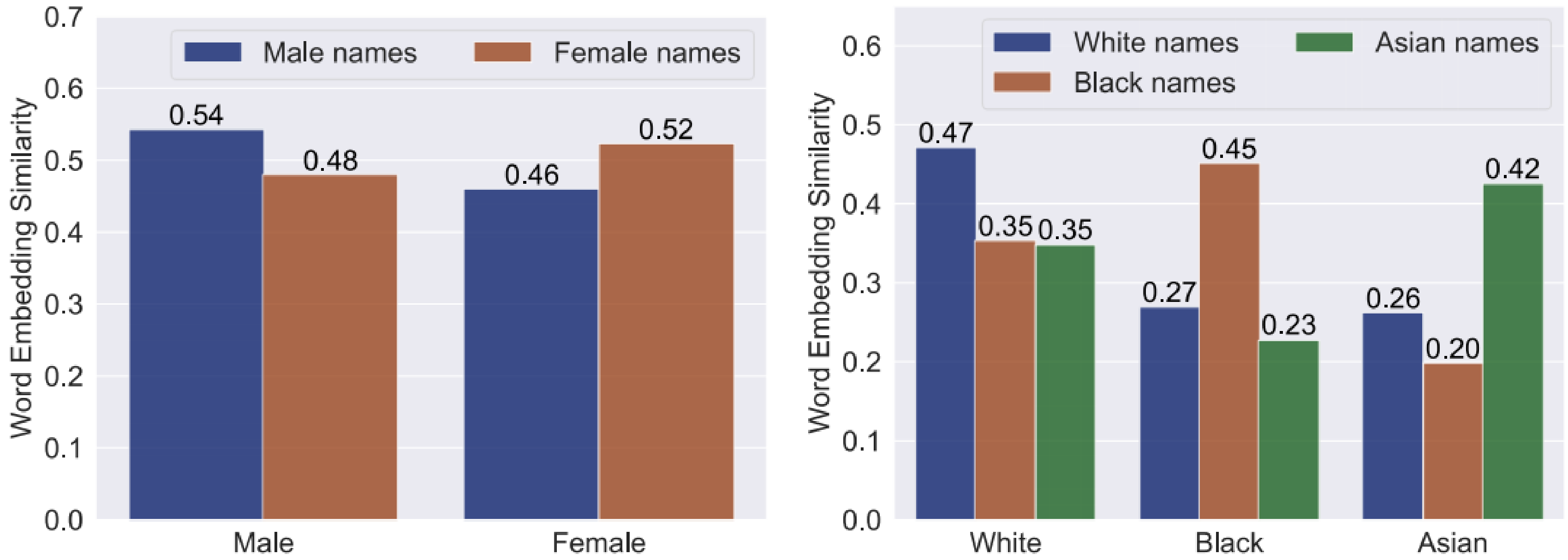
(a) Job Gender (b) Job Race (c) News Gender (d) News Race

Figure 2: The discriminatory behaviors (*i.e.*, topic distribution $P(L_K(s))$) against certain topics of LLMs under job and news domain for user names belonging to different Gender and Race groups.

“LLMs deliver more political but less art news to black users…”
“As for job recommendations, LLMs tend to recommend more service-related but less educational jobs to black users…”
“LLMs are likely to give more business and educational jobs to White and Asian users”

https://arxiv.org/pdf/2311.07054

Are your LLM-based IR biased or unfair?



(a) Gender

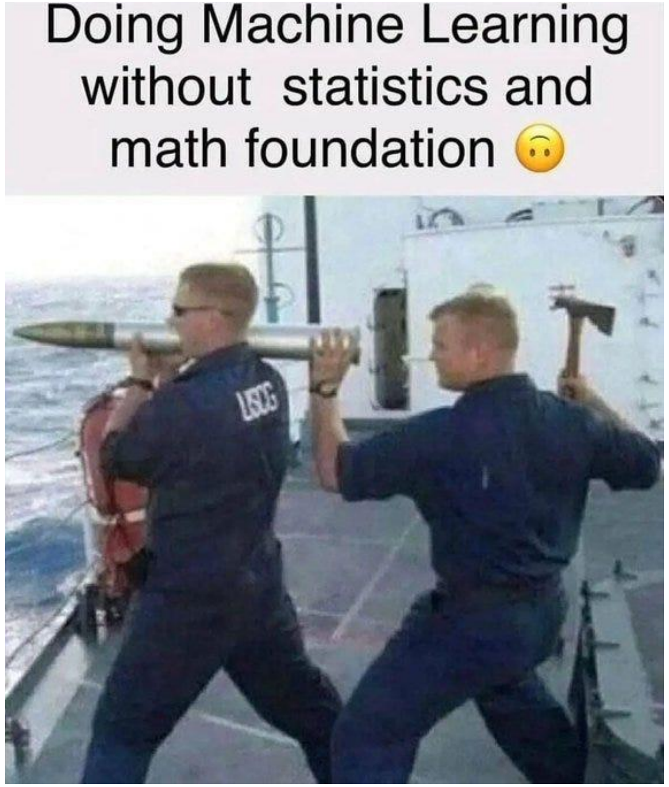
(b) Race

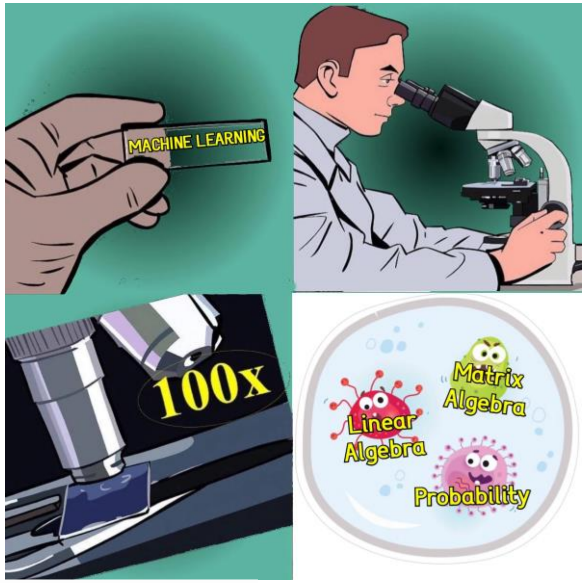
Figure 6: Word embeddings similarities between user names and sensitive attribute words.

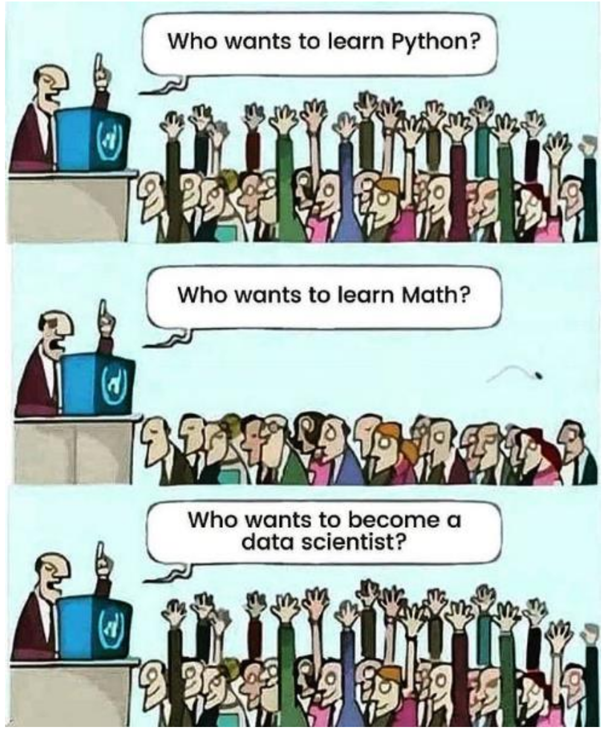
https://arxiv.org/pdf/2311.07054

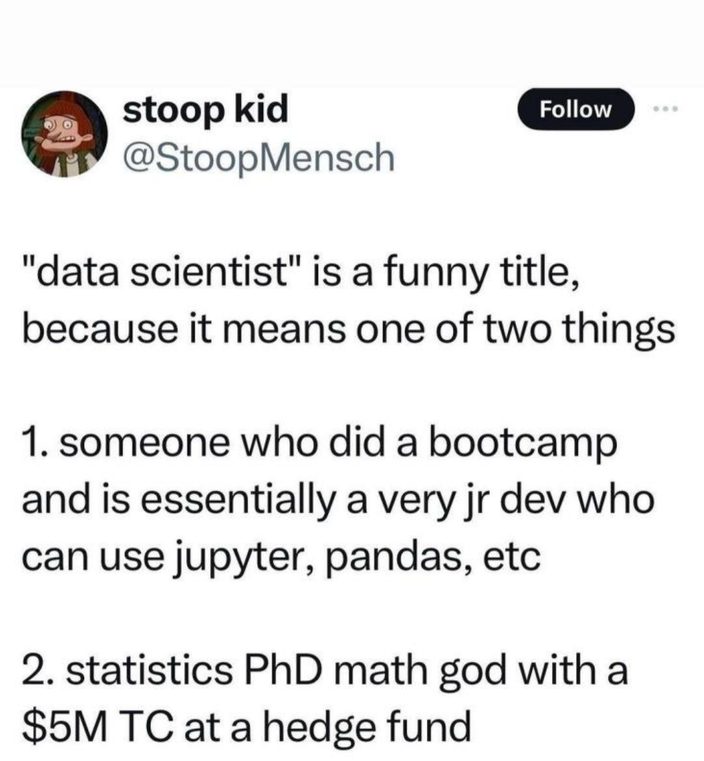
Few Last Words

Jangan tertipu (baca: "FOMO") dengan *buzz words*: AI, LLMs, Deep Learning, GenAI, Data Science, Big Data, …


Doing Machine Learning without statistics and math foundation 🙃
USCG




Who wants to learn Python?
Who wants to learn Math?
Who wants to become a data scientist?

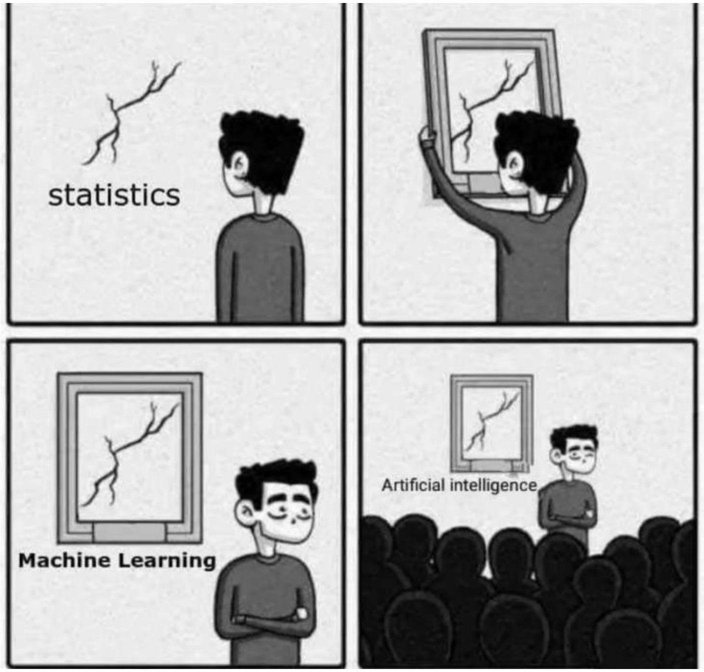


stoop kid
@StoopMensch

"data scientist" is a funny title, because it means one of two things

1. someone who did a bootcamp and is essentially a very jr dev who can use jupyter, pandas, etc

2. statistics PhD math god with a $5M TC at a hedge fund



Tanpa malu, meme-meme ini diperoleh dari Google Search.

MABA CS UI

```
>>> print: "hello, world!"
>>> hello, world!
```

First semester learning computer science

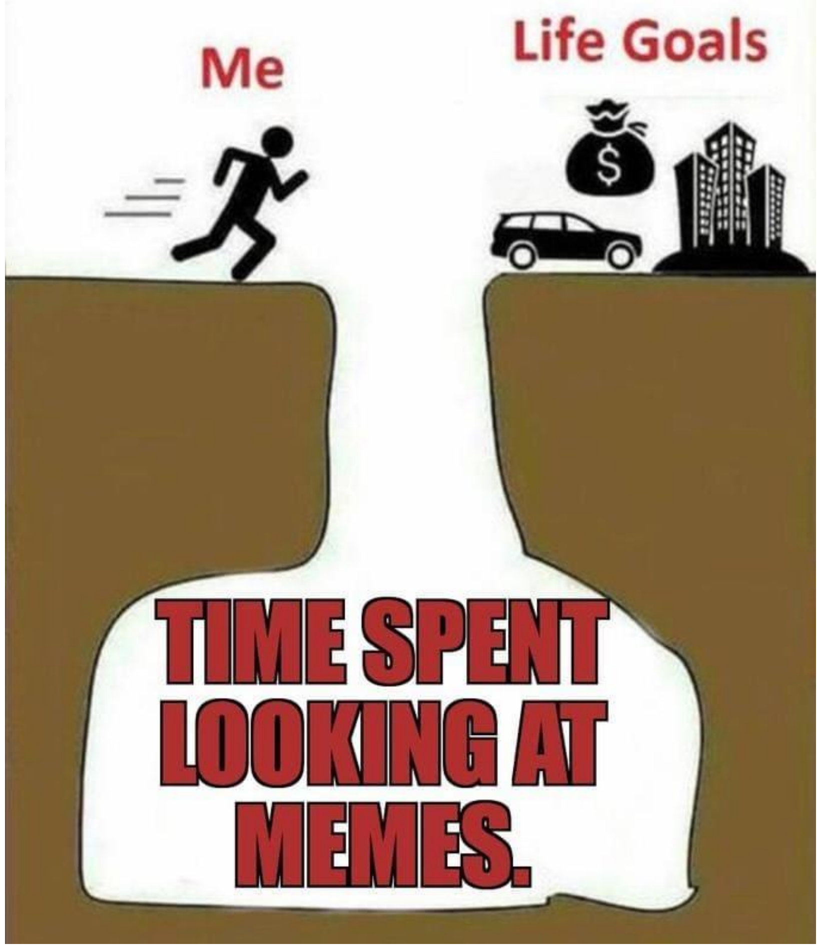
LULUS

Tanpa malu, meme-meme ini diperoleh
dari Google Search.

Lupakan AI, Lupakan LLMs, Lupakan Deep Learning, dan semuanya ……… untuk sementara waktu.

Mari sejenak renungkan pertanyaan mendasar: **untuk apa Anda kuliah di Fasilkom UI**? Apa rencana 5-10 tahun kedepan? …

Life should be goal-directed (and yes, this is only my opinion; you may disagree with this).



Dan saya sudah **buang-buang waktu sekitar 1.5 jam** untuk cari meme buat presentasi kuliah hari ini.

Tanpa malu, meme-meme ini diperoleh dari Google Search.