Portfolio

Jonghyun (Jong) Song
Ph.D. Student @ Seoul National University

Outlines

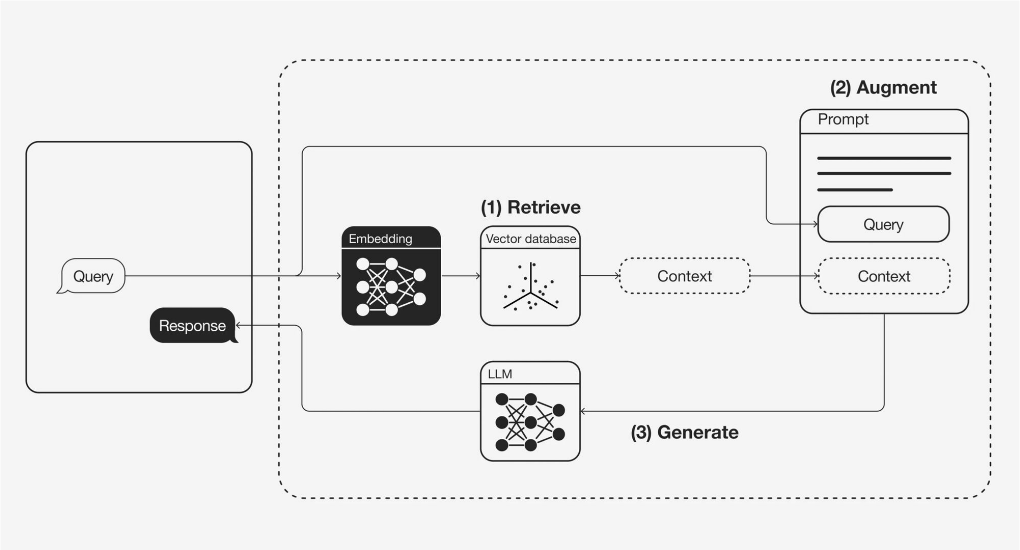
1. Injecting document-document interaction to information retrievers
- Motivations
- Project 1: Jointly Comparing Multiple Candidates (CMC; EMNLP main)
- Project 2: Beam Document Search for Complex QA (In progress)

2. Beam Document Search for Complex QA
- Motivations
- Project 1: Multi-modal Multi-view Patent Search Engine (1 Minister's award)
- Project 2: Redefining Information Extraction from Visually Rich Documents as Token Classification (1 IJCAI competition award)
- Project 3: Enhancing Performance of LLM for Understanding Documents through Various Markup Languages (In progress)

1. Injecting Document-document Interaction to Information Retrievers

Motivations

We need Good Retrievers!



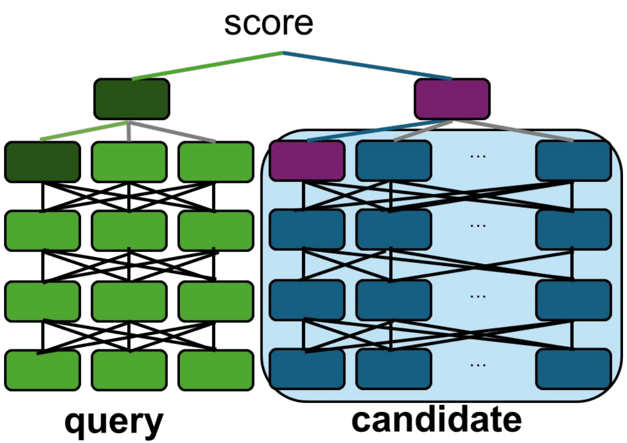
RAG pipeline

LLMs cannot answer complex questions in real-world contexts on their own.

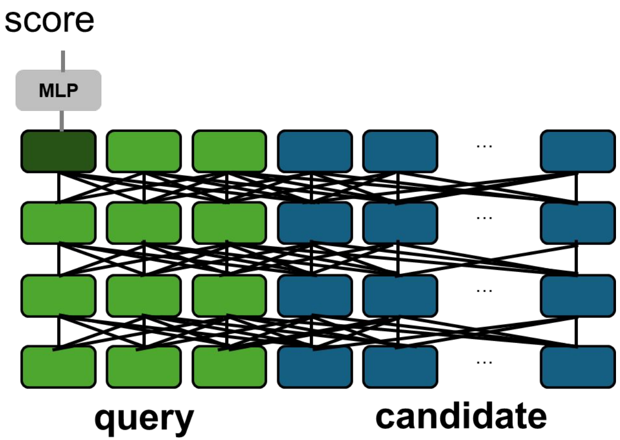
We need an **effective retriever**

We need Good Retrievers!

- Retrievers = compatibility evaluators for (query, document) pairs



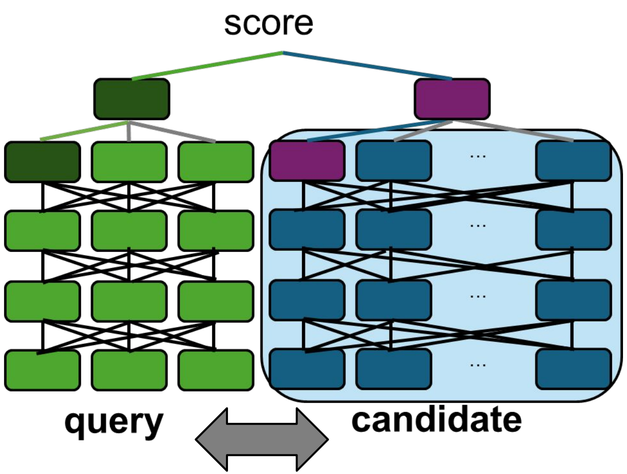
Bi-encoders
(BE; Retriever)



Cross-encoder
(CE; Reranker)

We need Good Retrievers!

- Retrievers = compatibility evaluators for (query, document) pairs



Bi-encoders

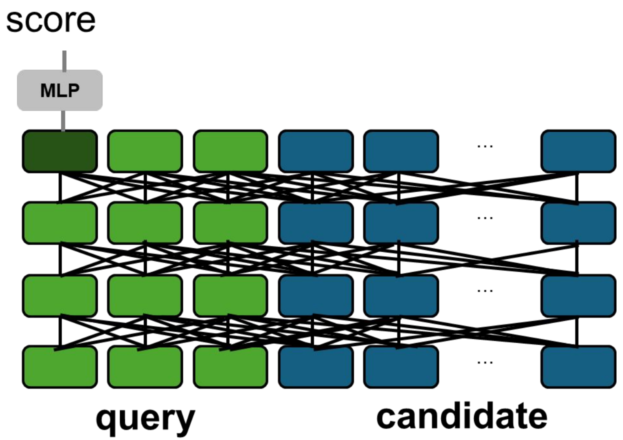
encode the query and candidates independently, then retrieve the **K** nearest candidates

(+) Efficient for large search spaces (MIPS)
(-) Inaccurate; May miss gold candidates

No token level interaction!

We need Good Retrievers!

- Retrievers = compatibility evaluators for (query, document) pairs



Cross-encoders

encode concatenated text and directly output a score for the final prediction.

(+) Candidates are examined more carefully
(-) Expensive and limited scalability

Cannot Search over large spaces!

We need Good Retrievers!

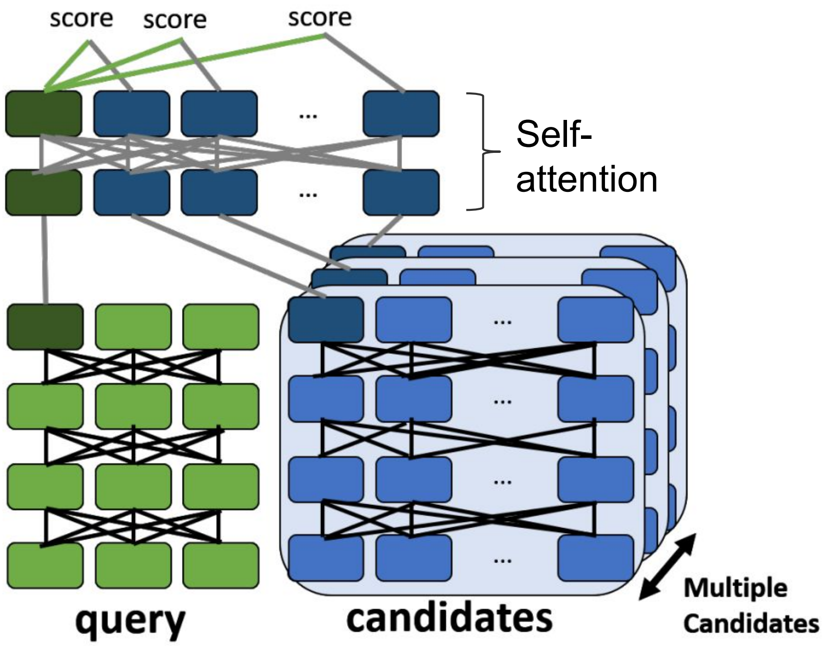
- Modeling efficient **query-document interaction** is important for retriever systems
  - **Fine-grained interaction** (i.e., cross-encoder) is accurate but computationally expensive
  - Coarse interaction (i.e., bi-encoder) is fast but less accurate

- Late interaction models are proposed to find a sweet spot
  - ColBERTv2(K Santhanam et al., 2021), Poly-encoder(S. Humeau et al., 2019), MixEncoder (Y. Yang et al., 2023)...

Project 1: Jointly Comparing Multiple Candidates (CMC)

1 EMNLP main paper accepted/ Spotlight paper talk at ACL workshop

Comparing Multiple Candidates (EMNLP 2024 Main)



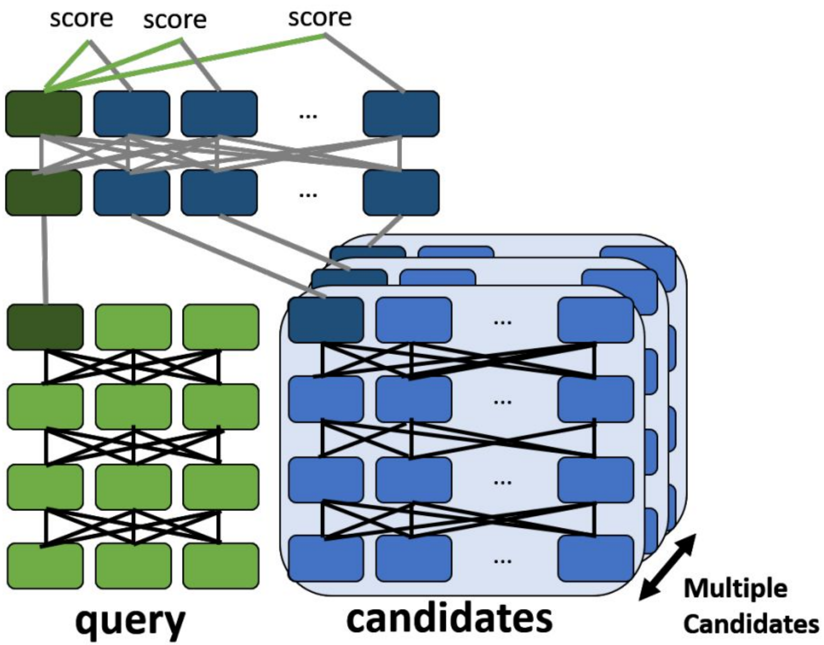
We are presenting **Comparing Multiple Candidates (CMC)** which

- condenses information to **single vector embedding** (~Bi-encoder), and uses **joint attention on query and multiple candidates** (~Cross-encoder)
- uses both **query-document** and **document-document** interaction via self-attention layer

J. Song et al., *Comparing Neighbors Together Makes it Easy: Jointly Comparing Multiple Candidates for Efficient and Effective Retrieval,* In EMNLP Main Track (Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing), 2024 / Spotlight Talk at 9th Workshop on Representation Learning for NLP in ACL 2024

Detail

Comparing Multiple Candidates (EMNLP 2024 Main)



(d) Comparing Multiple Candidates (CMC;Ours)

How CMC works?
Given candidates from the first-stage retriever (e.g., bi-encoder),

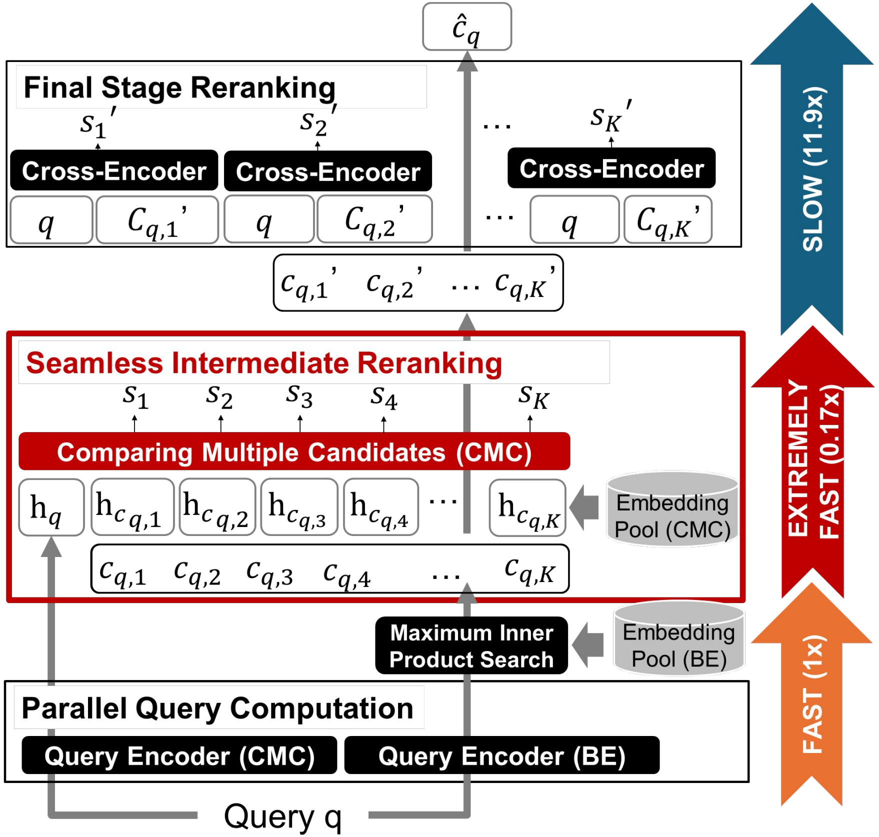
1. Query ($\mathbf{h}_q^{sent}$) and multiple candidates ($\mathrm{h}_{c_{q,j}}^{sent}$) are **pre-computed** like bi-encoders
2. The self-attention layer **jointly processes concatenated embeddings** of a query and all candidates

$$[\mathrm{h}_q^{CMC}; \mathrm{h}_{c_{q,1}}^{CMC}; \ldots; \mathrm{h}_{c_{q,K}}^{CMC}] = \mathrm{SelfAttn}([\mathrm{h}_q^{sent}; \mathrm{h}_{c_{q,1}}^{sent}; \ldots; \mathrm{h}_{c_{q,K}}^{sent}])$$

3. Final prediction is **calculated via dot products**

$$\hat{c}_q = \mathbf{argmax}_{c_{q,j} \in C_q} \mathrm{h}_q^{CMC} \cdot (\mathrm{h}_{c_{q,j}}^{CMC})^T$$

Comparing Multiple Candidates (EMNLP 2024 Main)

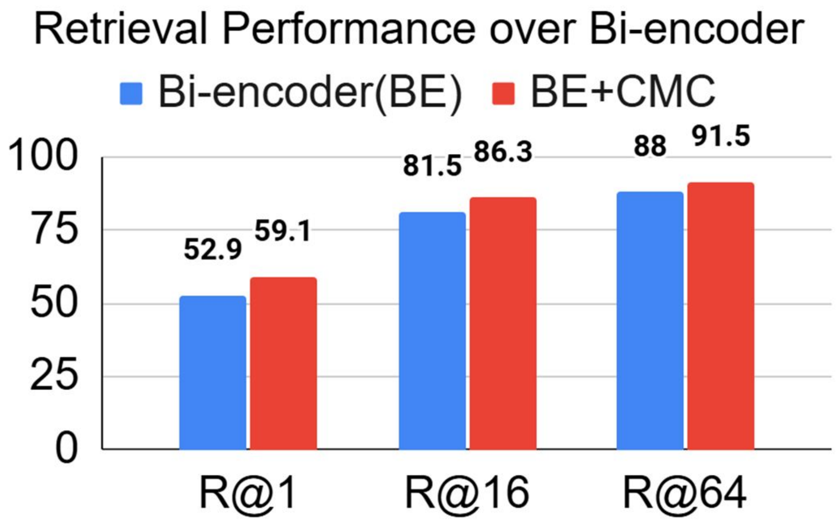


- CMC as the ***seamless intermediate reranker*** (BE-**CMC**-CE)
  - enhance retrieval performance with negligible extra latency
  - prevent error propagation from retriever
- CMC as a fast and effective ***final stage reranker*** (BE-**CMC**)
  - CMC can serve as the final reranker under time constraints

Comparing Multiple Candidates (EMNLP 2024 Main)

Performance as a Intermediate Reranker (BE-**CMC**-CE)

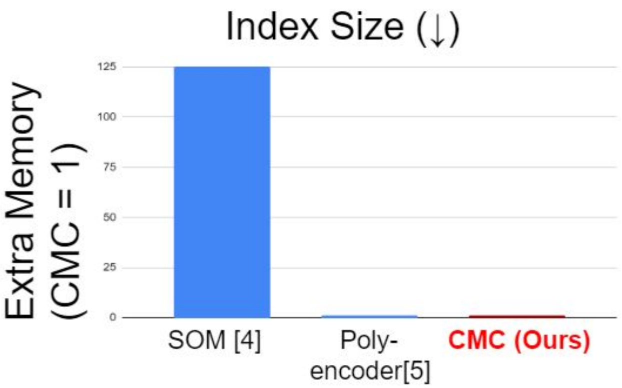
- CMC significantly improves **bi-encoder** Recall@K (+4.8p, 3.5p for R@16, R@64) at **a marginal extra speed (+0.07x)**

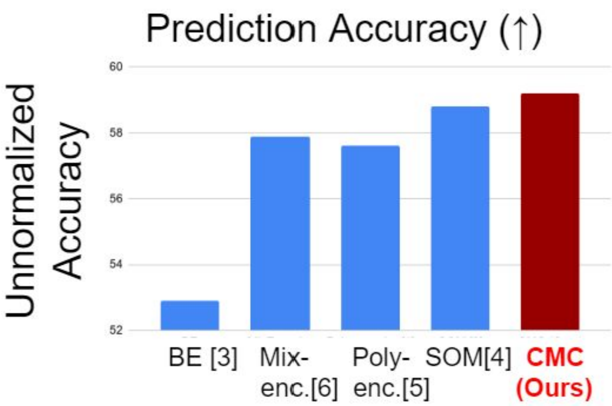


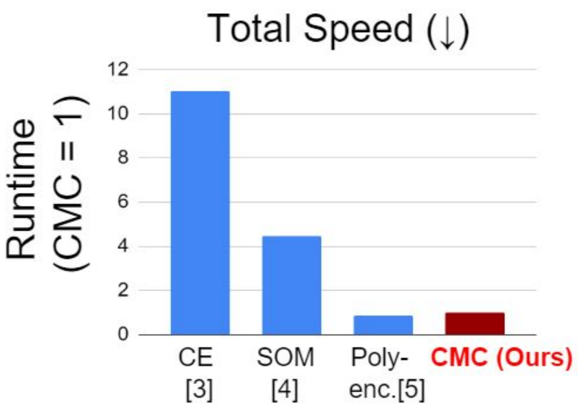
Comparing Multiple Candidates (EMNLP 2024 Main)

Performance as a Final-stage Reranker (BE-**CMC**)

- CMC shows robust performance over 4 datasets with 3 tasks
- CMC is 11x faster than cross-encoders and requires 125x less index size than Sum-of-max







Project 2: Beam Document Search for Complex QA

In progress

Ongoing Projects: Retriever for Complex QA

CMC can effectively replace token-level interaction of cross-encoder with **query-document and document-document and interaction with single vector embeddings**

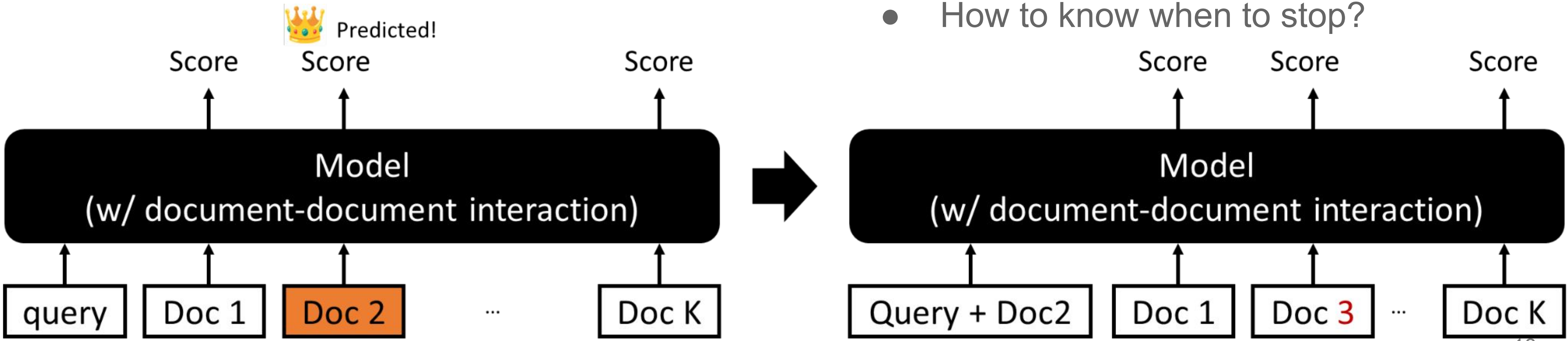
→ *What if we apply document-document interaction to* ***complex QA tasks****, where multiple documents are required to be retrieved?*

Ongoing Projects: Retriever for Complex QA

- Project: Injecting document interactions for *Document Set* Retrieval
- Task: Given query q, predict the set of documents D
- Method: Beam Document Search

Remaining Challenges

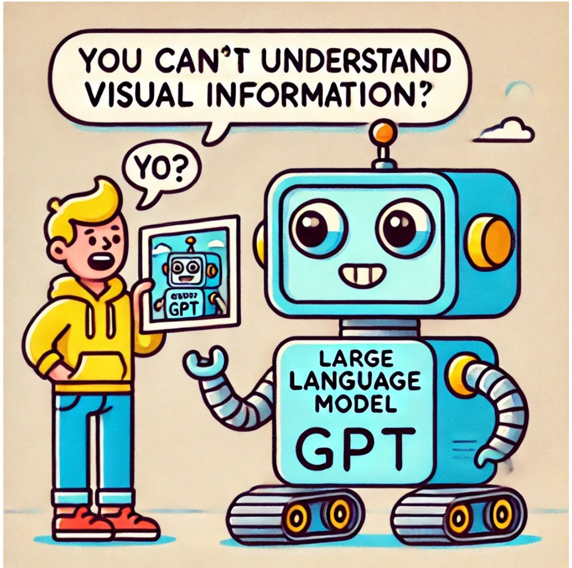
- How to efficiently consider document interaction in large search space?
- How to know when to stop?



2. Enhancing Performance of LLM for Understanding Document Images

Motivations

Can LLM understand complex document Images?

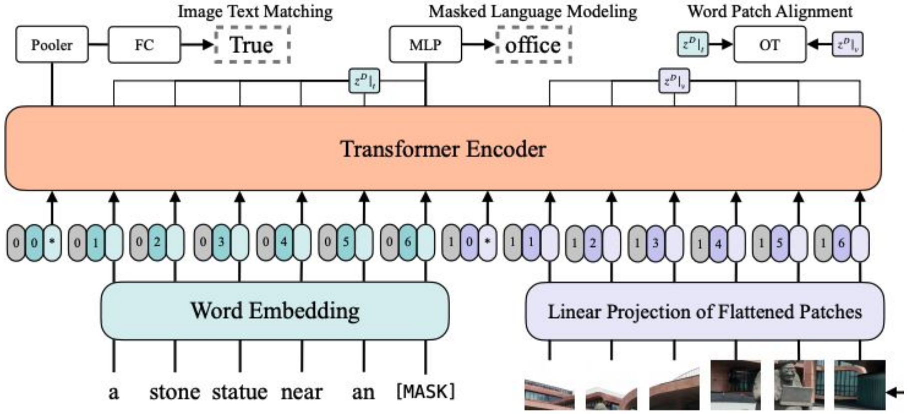


DALL-E generated

Research Question:

How can LLM (not LMM) understand or generate visual (or multi-modal) information?

Can LLM understand complex document Images?



LLM
(Large Language Models)

(+) Superior Reasoning Capability

(-) Not understand multi-modal information

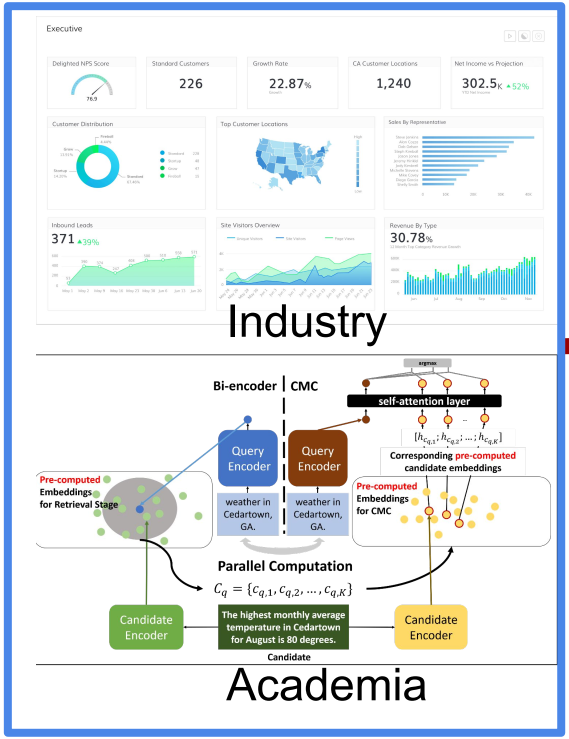
LMM
(Large Multi-modal Models)

(+) Access to visual information

(-) Inferior reasoning capability

(-) difficulty understanding images with text

Can LLM understand complex document Images?



Document

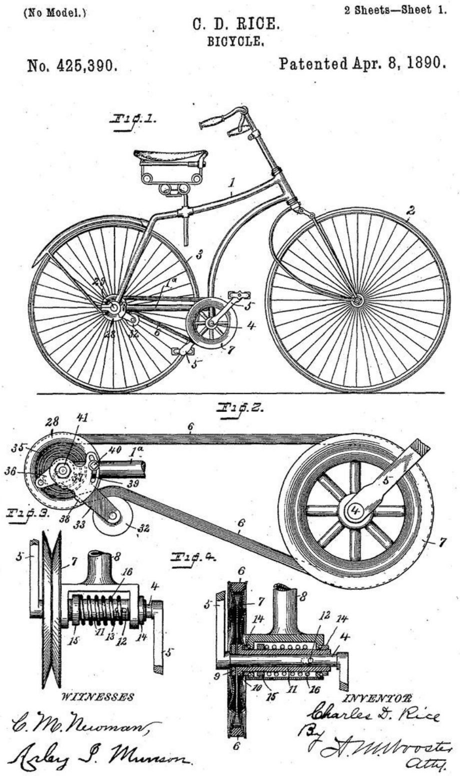
How can we represent these images effectively?

Project 1: Multi-modal Multi-view Patent Search Engine

1 Minister's Award @ Korea-Data Science Hackathon

Introduction

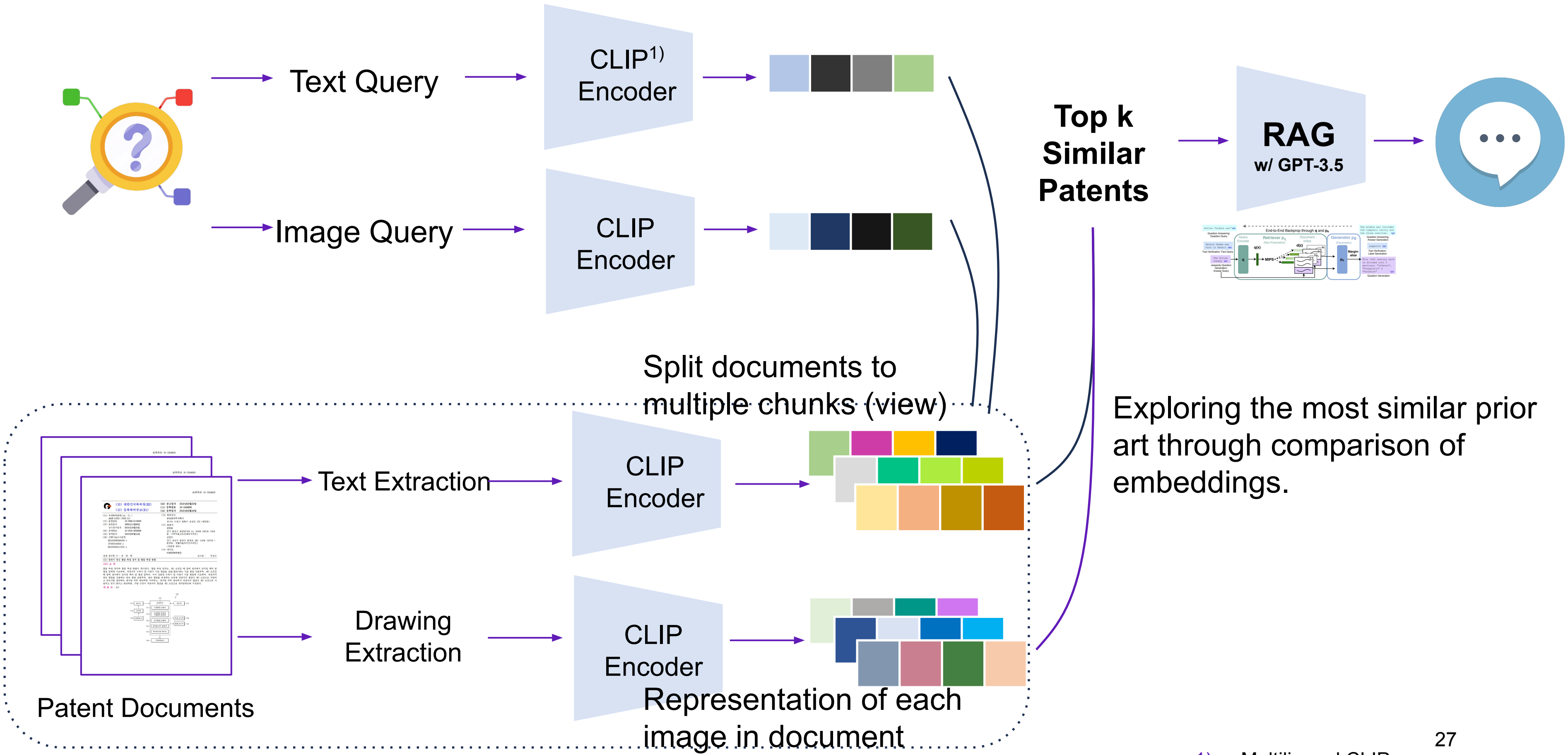
- Prior work search for patents search is **important** but **difficult**
  - Patent description has **long-context** and **mixed modality**
  - Patent attorney often uses techniques to avoid specific keywords not to be expose their patents



Solutions: Multi-modal Multi-view Patent Search

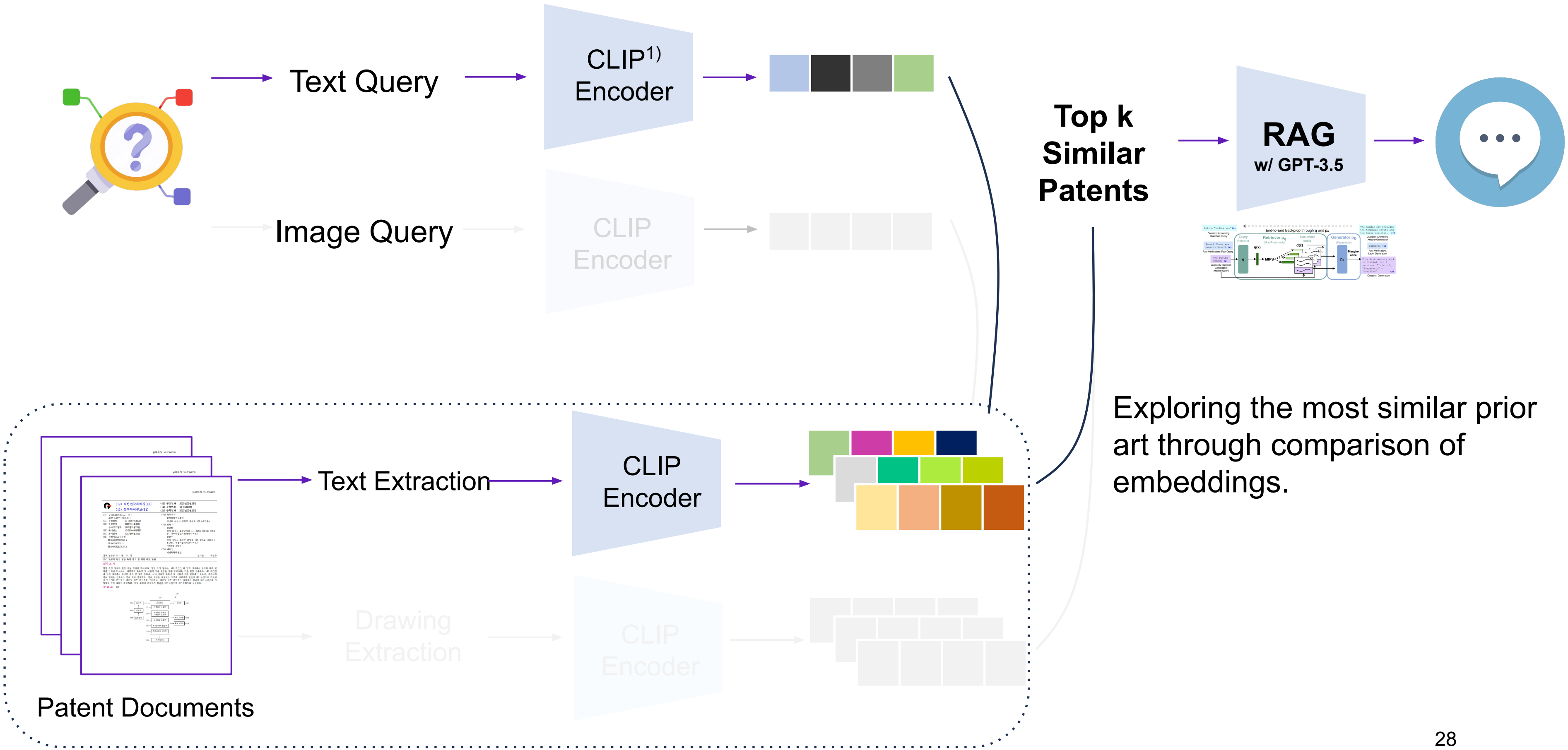
Multi-Modal	Using a multimodal model capable of handling **both text and images**, we will conduct a similarity search for patents
Multi-View	Dividing lengthy patent documents into multiple **chunks (i.e., views)**, embedding them, and analyzing embedding similarity to semantically search through patent documents regardless of their location
RAG (Retrieval-Augmented Generation)	Developing a chatbot-style UI that provides detailed answers to questions based on search results, rather than simple search result returns

Model Architecture - CLIP Embedding



1) Multilingual CLIP

Model Architecture (Fast Mode)

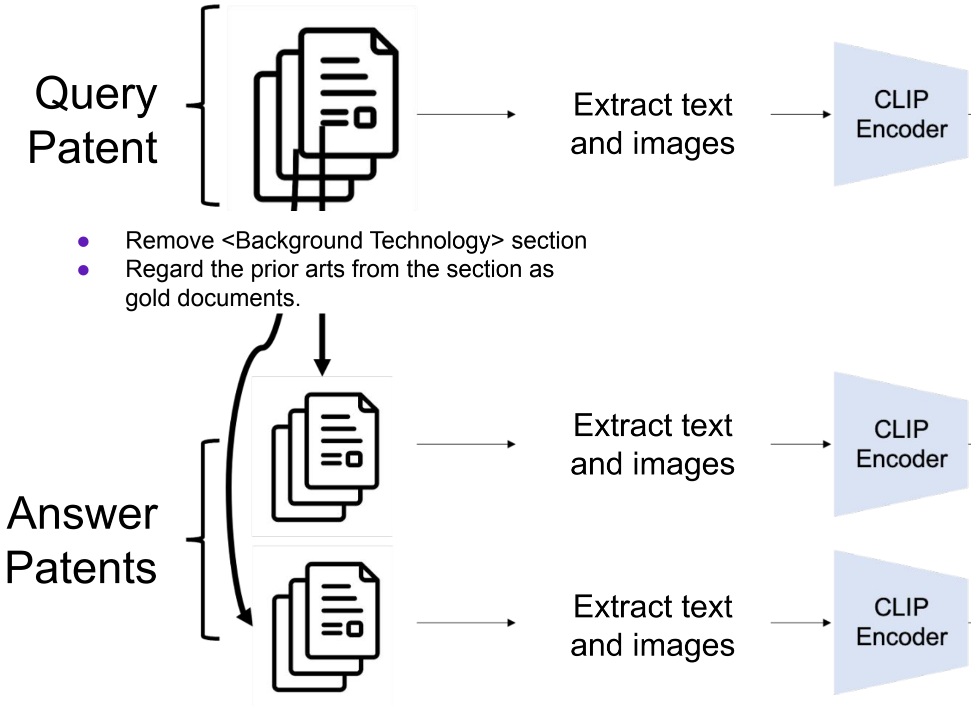


Exploring the most similar prior art through comparison of embeddings.

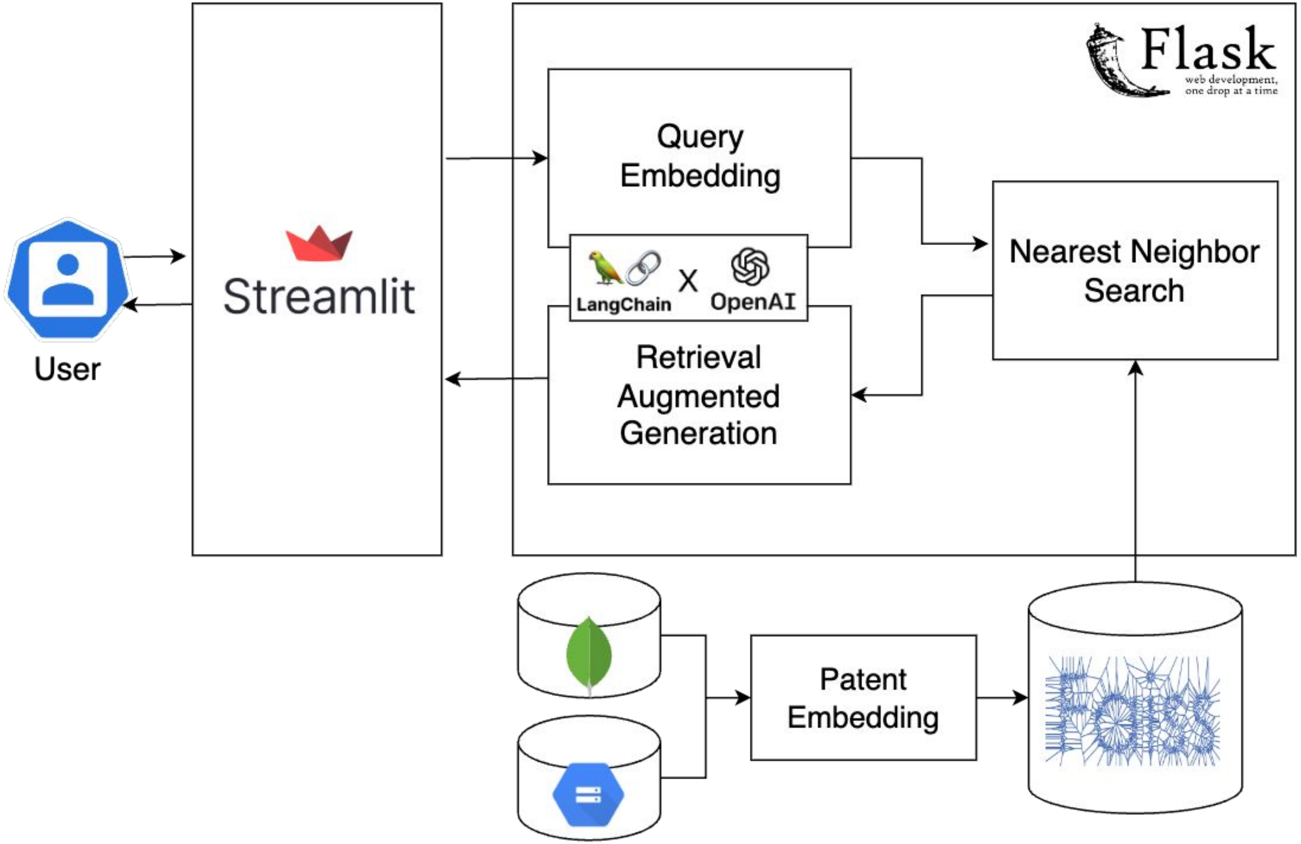
1) Multilingual CLIP is used.

Training Strategies

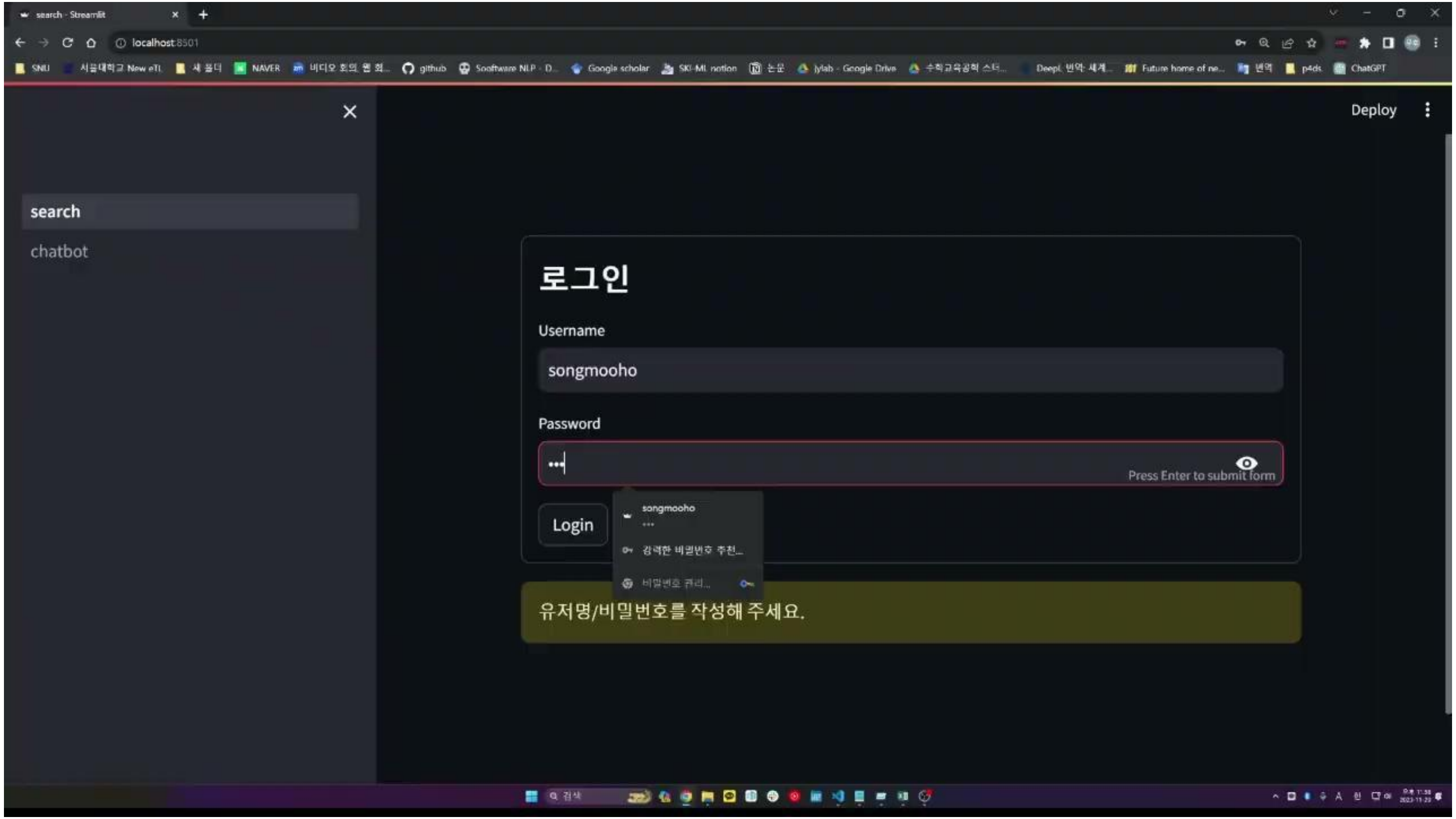
- Self-supervised Training
  - Some patents have designated **‘prior arts’** in the section ‘Background Technology’
  - Regard this as similar items in contrastive learning



System architecture



Demonstration (Korean)



Project 2: Redefining Information Extraction from Visually Rich Documents as Token Classification

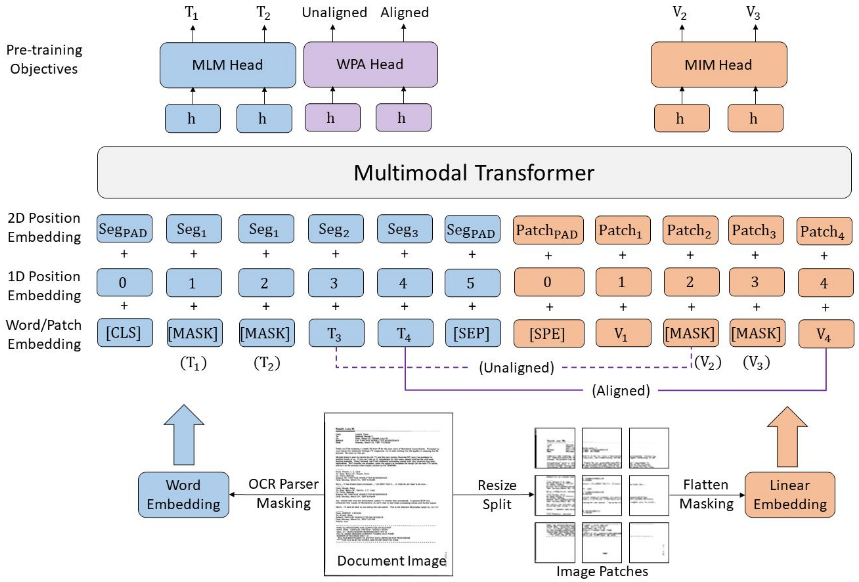
2nd Place @ IJCAI 2024 Competition on visually rich document understanding

Background: Form-NLU Datasets

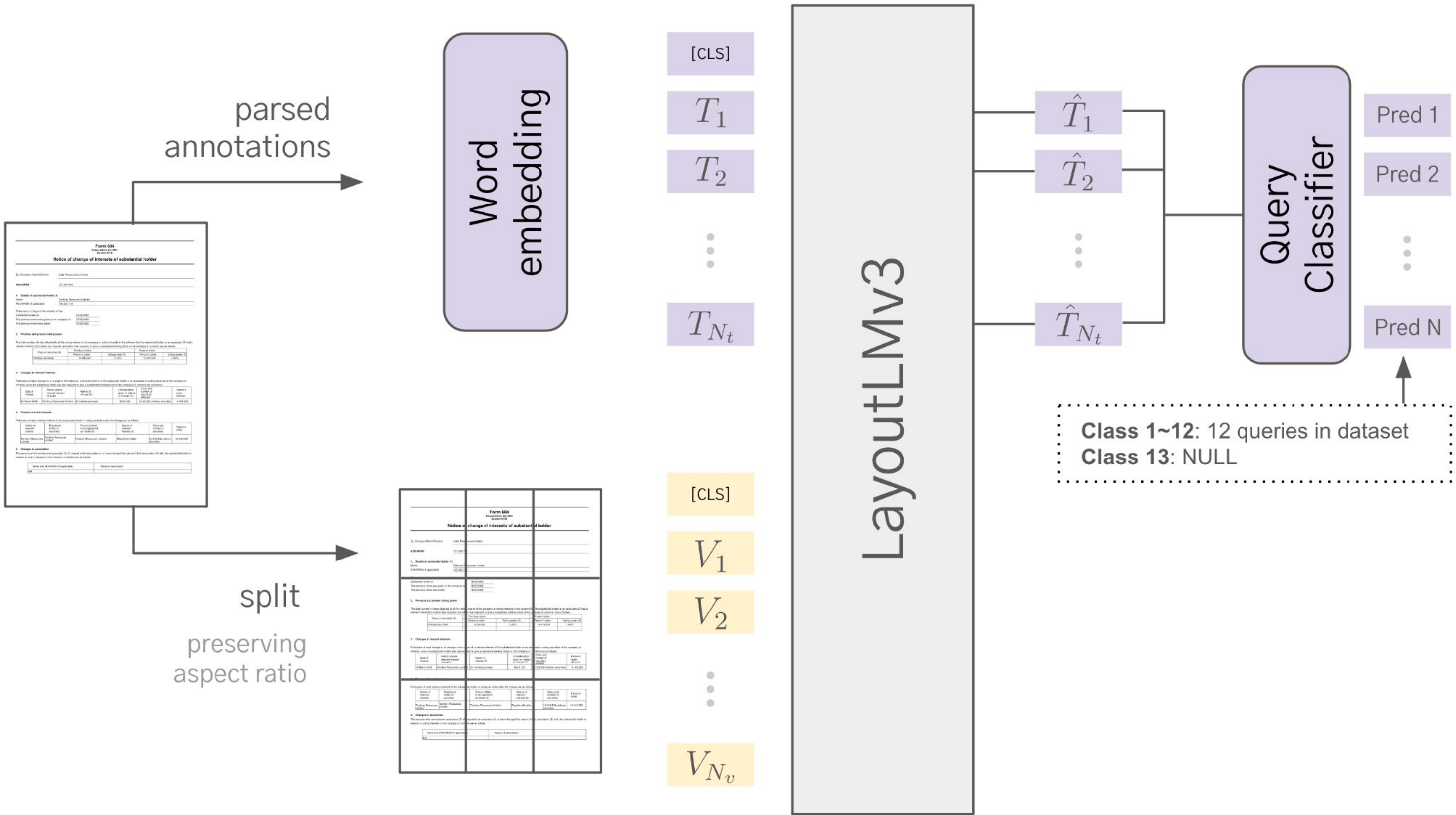
- Queries (keys) related to form designers' intentions **are limited**:
  - Only 12 queries are presented
- Includes **digital, printed, and handwritten images**
- Various meta information of ROIs is presented
  - e.g., text, bounding box coordinates, text/visual feature, etc.
- Document may have **no values related to keys** (i.e., NIL prediction)

Background: LayoutLMv3 (Huang et al., 2023)

- **Multimodal transformer** for document understanding
- Pre-trained objectives:
  - Masked Language Modeling
  - Masked Image Modeling
  - Word-Patch Alignment
- Pre-trained dataset:
  - IIT-CDIP Test Collection 1.0 - a **large-scale scanned document image dataset**

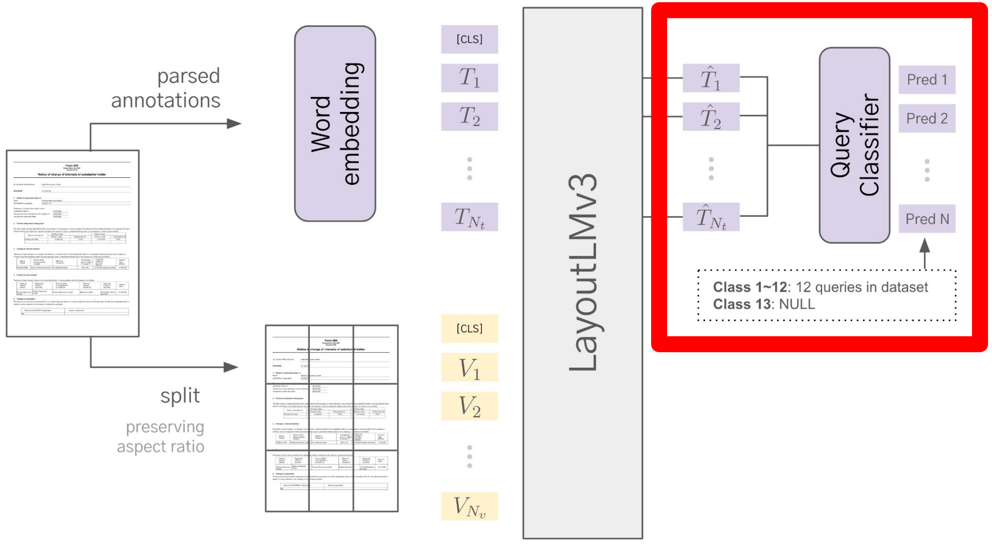


Solution: Information Extraction as Token Classification



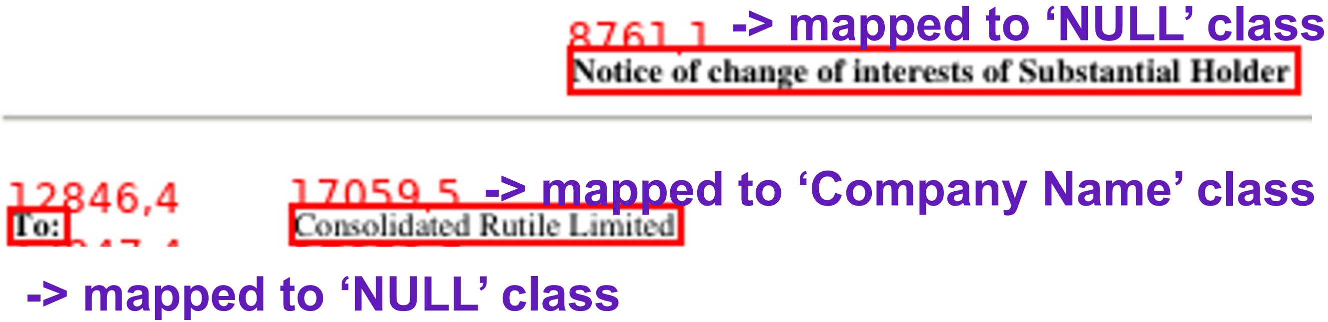
Solution: Information Extraction as Token Classification

- Redefining Information Extraction as **Token Classification**
  - As # of queries is limited to 12, we define the problem as token classification task with **13 classes** (12 queries + 1 for NULL)



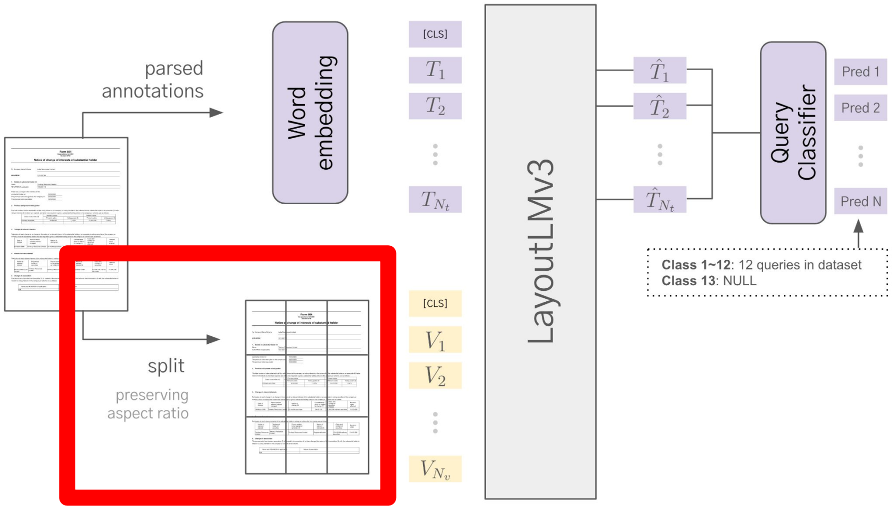
Solution: Information Extraction as Token Classification

- Redefining Information Extraction as **Token Classification**
  - As # of queries is limited to 12, we define the problem as token classification task with **13 classes** (12 queries + 1 for NULL)



Solution: Information Extraction as Token Classification

- **Preserving aspect ratios** of document images
  - Document images include text, which might be affected by the aspect ratio of the images.
  - **Retaining original aspect ratios** as much as possible (600 by 800)



Result

- Our model shows **robust performance** on both the public and private datasets
- Maintaining a resolution close to the original aspect ratio (600 by 800) significantly improves performance on the public dataset.

Model	Steps	Resolution	public	private
LayoutLMv3	10K	(224, 224)	96.55	97.75
		(600, 800)	97.60	**97.93**
	100K	(224, 224)	96.02	97.75
		(600, 800)	**97.77**	96.72

Failure Cases: Inference w/ GPT-3.5-turbo

- We prompted **text-only GPT-3.5-turbo** with text information of the objects
- Other techniques such as one-shot chain-of-thought prompting are also deployed

Prompt: Given objects from financial form. The answer can only be extracted from this list:

- global id: 18191, text: Form 604 Corporations Act 2001 Section 671B, center x_axis: 264.0, center y_axis: 41.0, width: 85.0, height: 44.0, category: 1
- global id: 18192, text: Notice of change of interests of substantial holder, center x_axis: 169.0, center y_axis: 89.0, width: 274.0, height: 16.0, category: 1
- global id: 18193, text: 1. Details of substantial holder (1), center x_axis: 73.0, center y_axis: 174.0, width: 123.0, height: 11.0, category: 2
- global id: 18194, text: 2. Previous and present voting power, center x_axis: 70.0, center y_axis: 309.0, width: 141.0, height: 13.0, category: 2

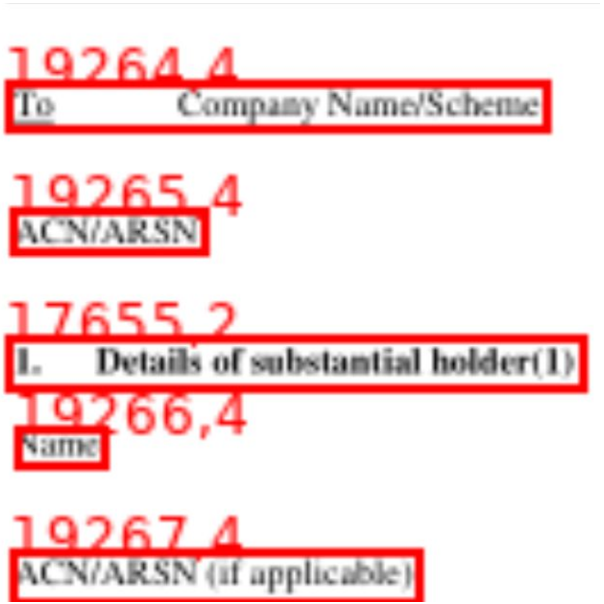
Failure Cases: Inference w/ GPT-3.5-turbo

- GPT-3.5 **does not** perform well, implying that the form-understanding capability of **text-only LLM is not well developed yet.**

Model	Steps	Resolution	public	private
LayoutLMv3	10K	(224, 224)	96.55	97.75
		(600, 800)	97.60	**97.93**
	100K	(224, 224)	96.02	97.75
		(600, 800)	**97.77**	96.72
GPT-3.5-turbo	-	-	31.77	38.28

Failure Cases: Inference w/ GPT-3.5-turbo

- GPT-3.5 **does not** perform well, implying that the form-understanding capability of **text-only LLM is not well developed yet.**
    - They do not understand information over multiple bounding boxes
        - e.g. GPT does not understand the key to ‘holder ACN/ARSN’ is bbox ‘19267’, not ‘19265’



Conclusion

- Fine-tuning a multi-modal transformer pre-trained with scanned documents (LayoutLMv3) shows robust performance on a diverse distribution of datasets (digital & printed).
  - Keeping aspect ratios similar to the original document is helpful in most cases
- Prompting document information to text-only LLM does not effectively solve the problem
  - Future work will explore the potential of LLMs (including Vision LLMs) for visually rich document understanding tasks.

Project 3: Enhancing Performance of LLM for Understanding Documents through Various Markup Languages

In Progress

Background

GPT did **NOT** understand plain text prompt well
in the competition

*What if prompt is given as **markup languages**?*

Research Questions

[RQ 1] Can LLMs better understand visually rich documents with OCR when they are expressed in **markup language** (e.g., HTML and XML etc.) rather than in **plain text with coordinates?**

[RQ 2] **Which data format is the most effective** for representing layout information of visually rich documents? i.e., which is the best format for LLM processing: HTML, XML or Markdown?

Baseline

- LMDX: Language Model-based Document Information Extraction and Localization (Perot et al., 2024)
  - This recent work focuses on using only simple text for LLM-based VRD understanding.
  - The document is represented in the format: <Text> XX|YY
  - Providing coordinate tokens led to a 14.98p↑ in F1 score on the VRDU-Ad-buy dataset."

Model architecture

- Overall pipeline
  - OCR is used to extract text and layout
  - OCR result is parsed to markup language (e.g., .html, .xml, and .md)
    - The markup language is expected to preserve both textual content and document layout better than plain text.
  - The structured markup language is processed by an LLM for the downstream task.