

# Retrieval **A**ugmented **G**eneration (RAG)

Super AI Engineer SS4

By Sarawoot Kongyoung

# Retrieval **A**ugmented **G**eneration (RAG)

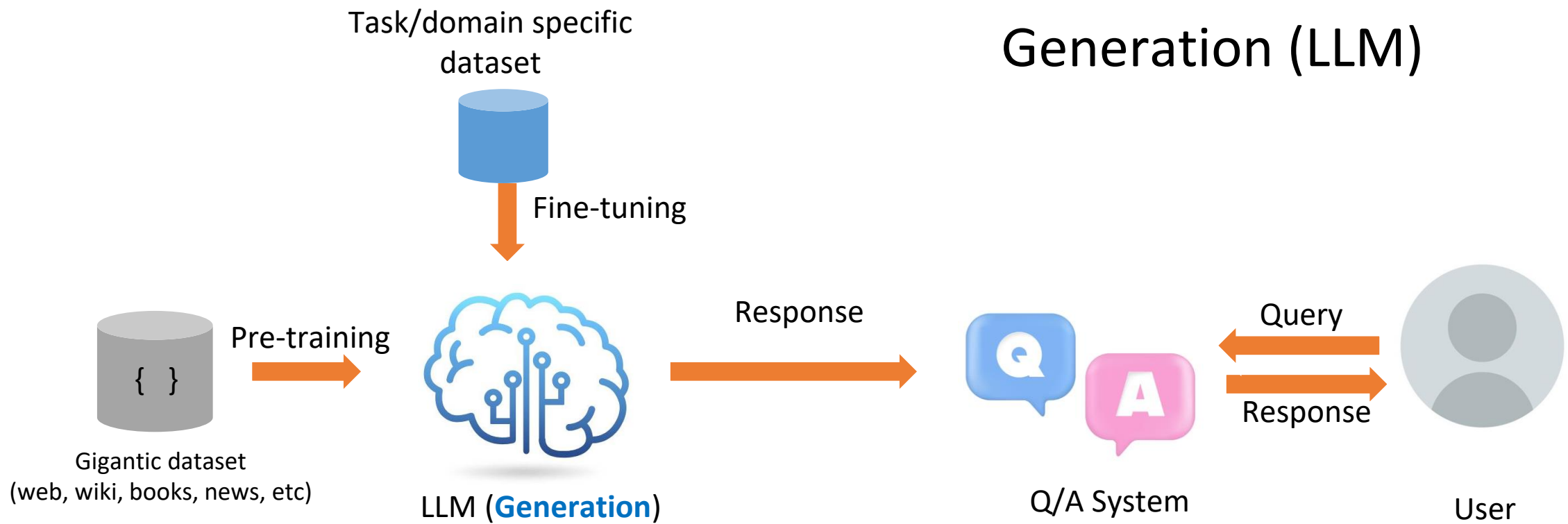


Retrieval

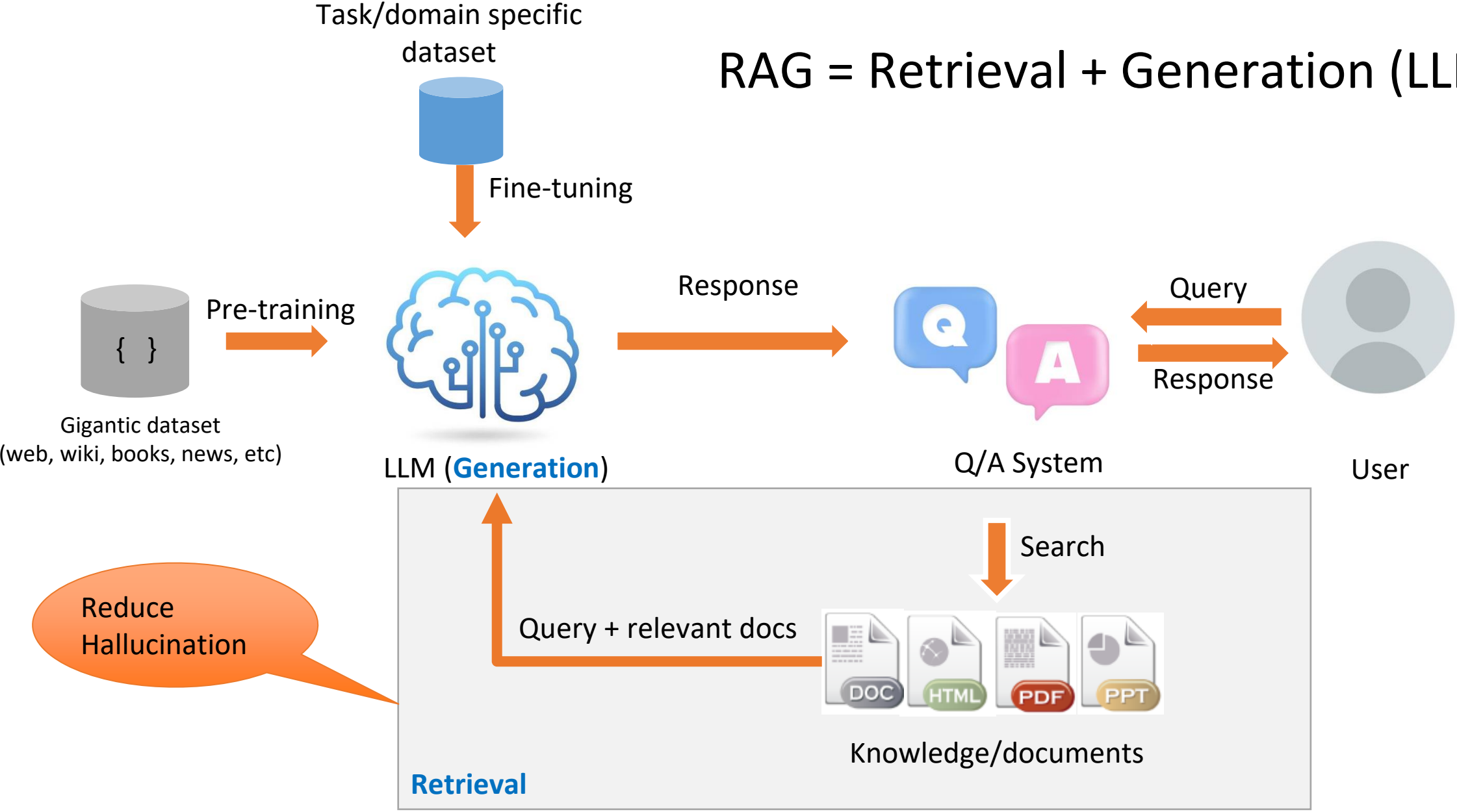
The diagram consists of a blue rectangular box on the left containing the word 'Retrieval', followed by an orange plus sign, and then a green rectangular box on the right containing the word 'Generation'.

+

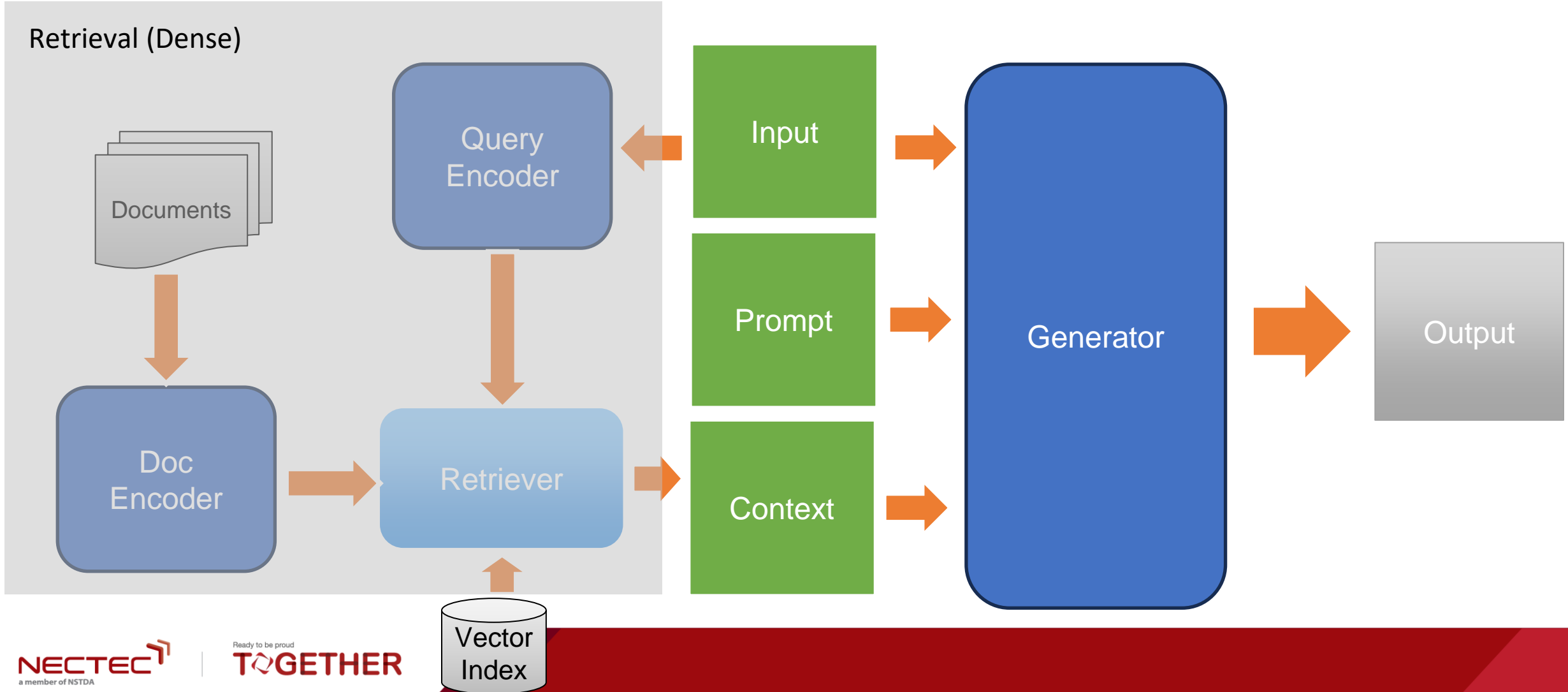
Generation



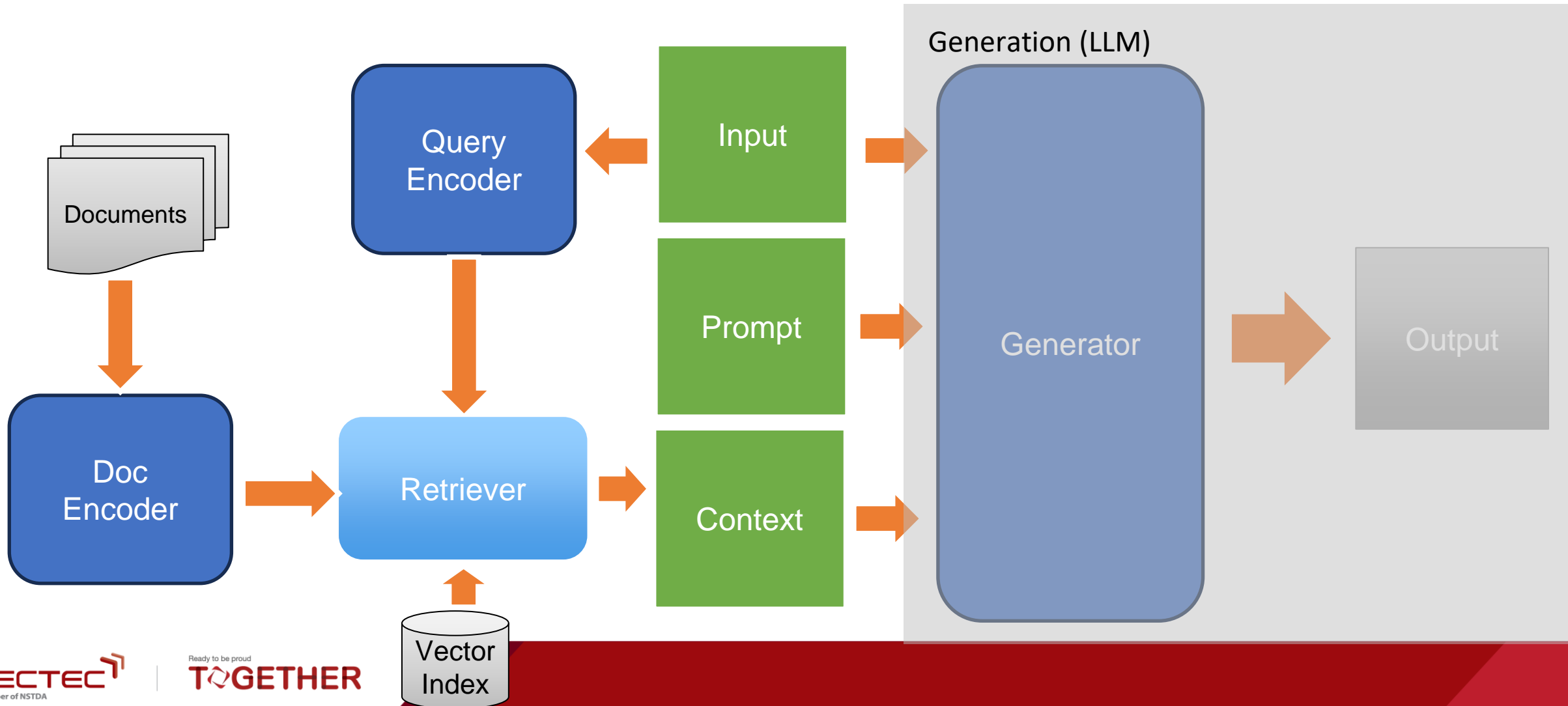
# RAG = Retrieval + Generation (LLM)



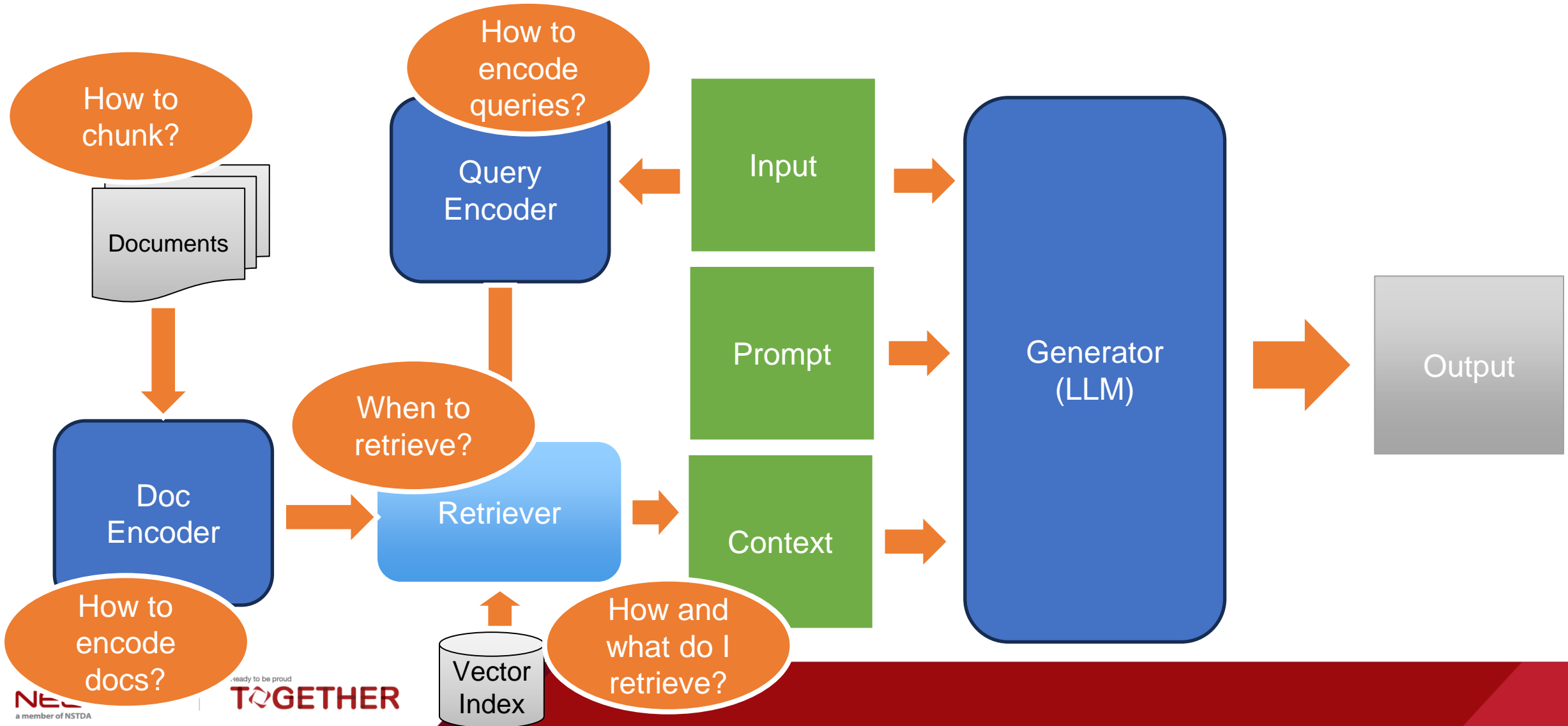
# RAG: Retrieval Augmented Generation



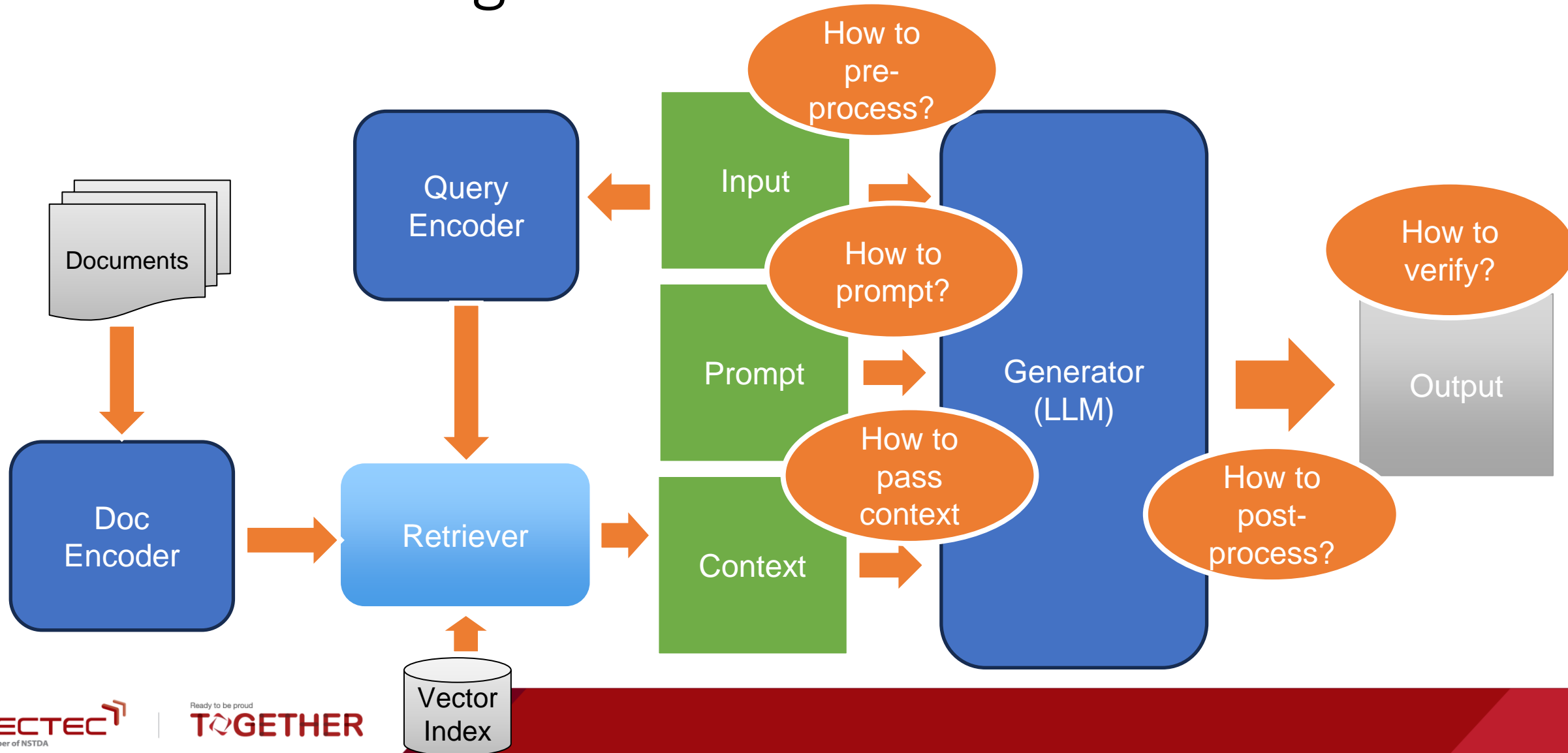
# RAG: Retrieval Augmented Generation



# RAG: Retrieval Augmented Generation

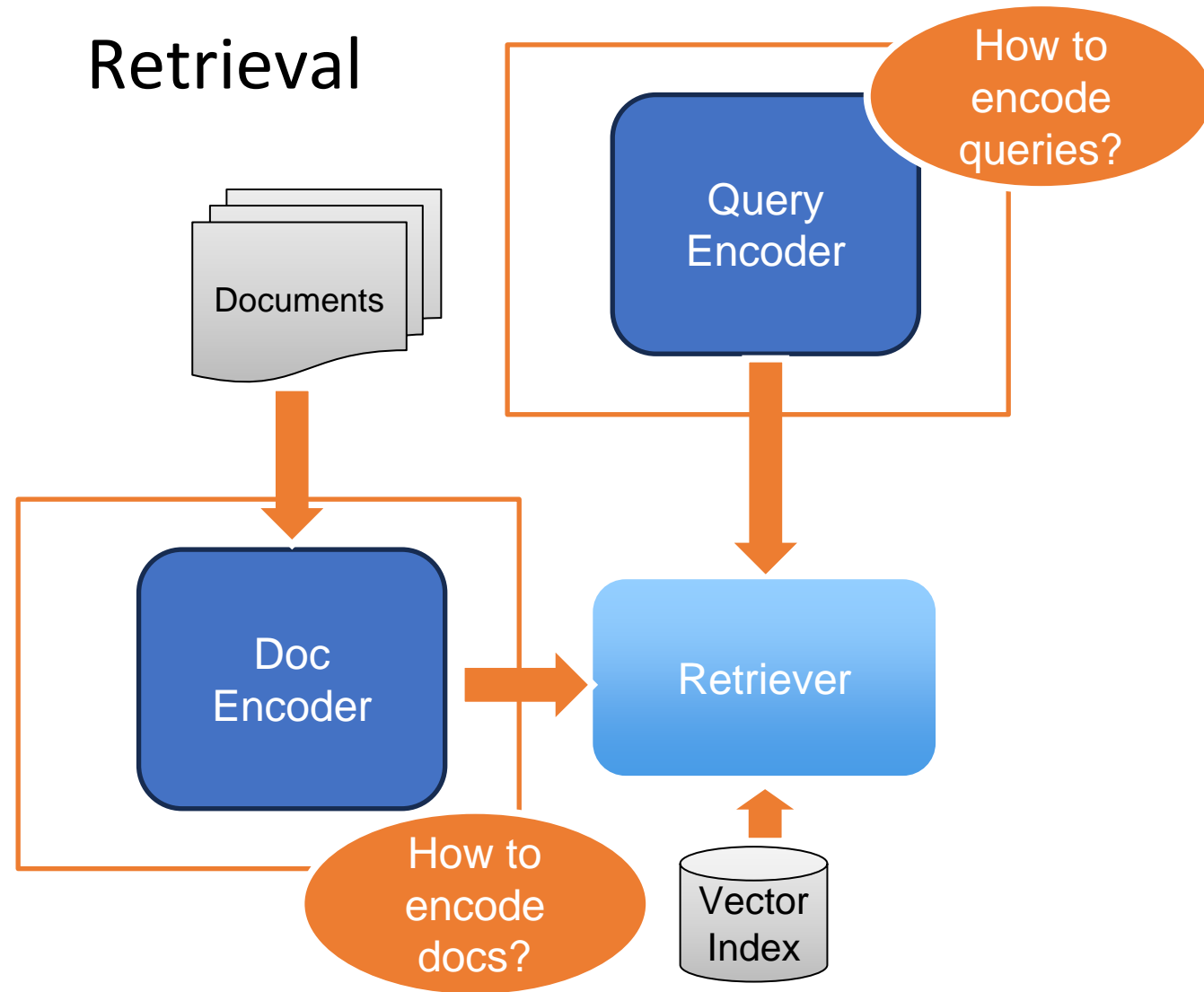


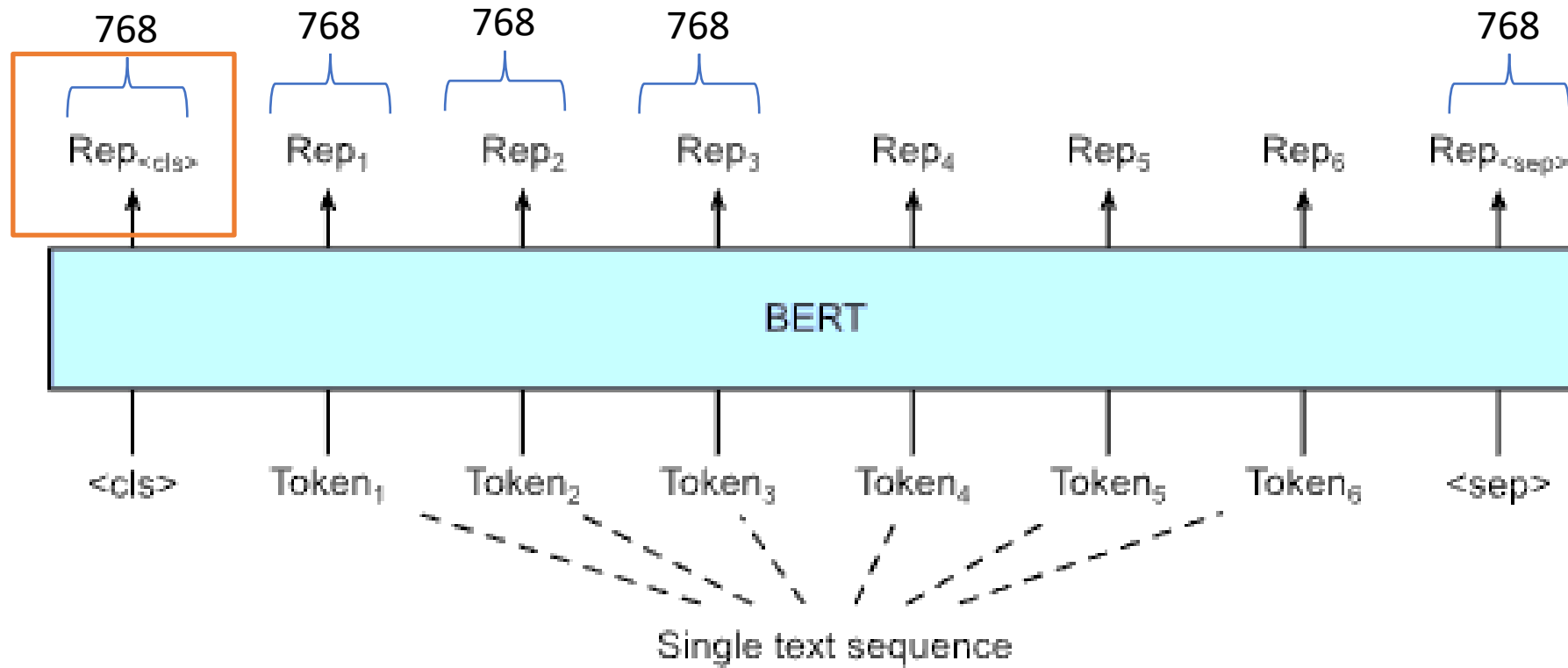
# RAG: Retrieval Augmented Generation





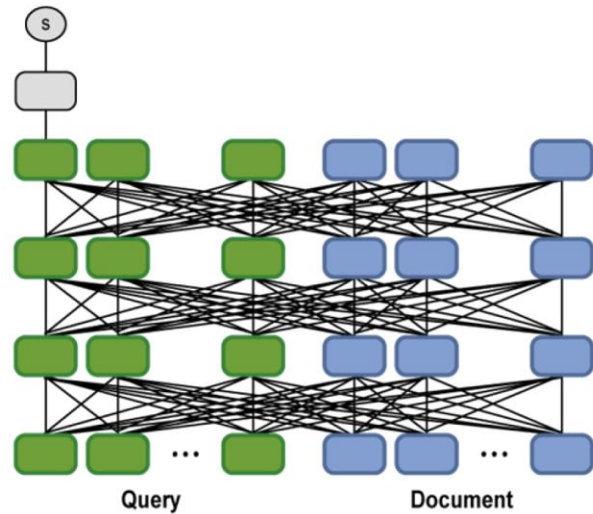
# Retrieval



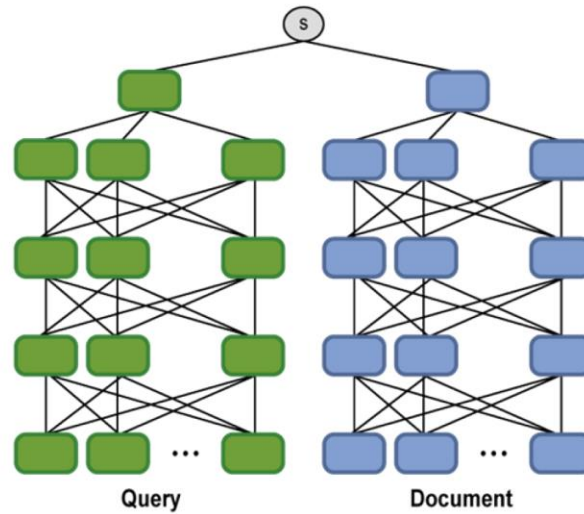


Source: [https://d2l.ai/chapter\\_natural-language-processing-applications/finetuning-bert.html](https://d2l.ai/chapter_natural-language-processing-applications/finetuning-bert.html)

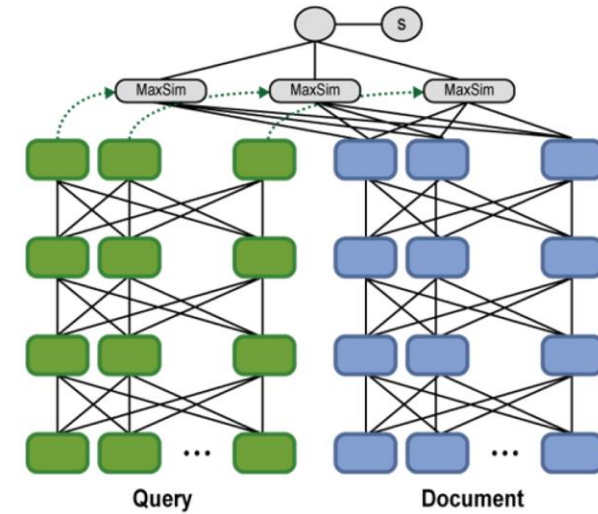
# Dense Retrieval



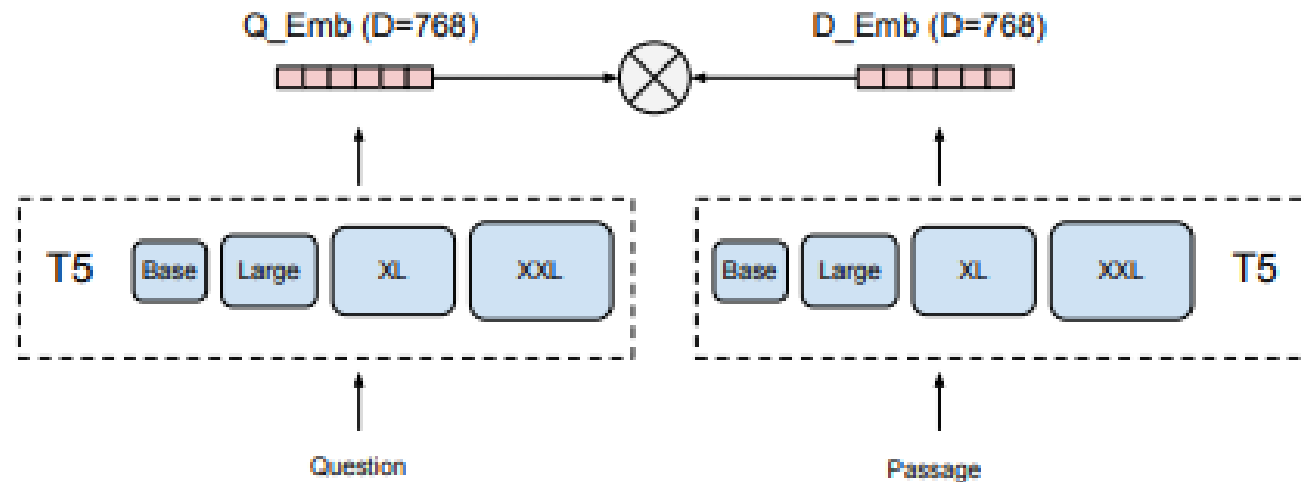
(a) Interaction-based retrieval  
(e.g., BERT)



(b) Representation-based retrieval  
(e.g., DPR)



(c) Representation-Interaction retrieval  
(e.g., ColBERT)



# Dense Retrieval for Thai?

- <https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark>

| Base Model  | R@1   | MRR@10 | Supervised? |
|---|-------|--------|-------------|
| <a href="#">simcse-model-distil-m-bert</a>            | 44.69 | 51.39  |             |
| <a href="#">simcse-model-m-bert-thai-cased</a>        | 45.09 | 52.37  |             |
| <a href="#">simcse-model-XLMR</a>                     | 58.06 | 64.72  |             |
| <a href="#">simcse-model-wangchanberta</a>            | 62.65 | 70.02  |             |
| <a href="#">simcse-model-phayathaibert</a>            | 71.43 | 78.16  |             |
| <a href="#">SCT-model-XLMR</a>                        | 49.28 | 58.62  |             |
| <a href="#">SCT-model-wangchanberta</a>               | 58.19 | 68.05  |             |
| <a href="#">SCT-model-phayathaibert</a>               | 63.43 | 71.73  |             |
| <a href="#">SCT-Distil-model-XLMR</a>                 | 56.36 | 65.18  |             |
| <a href="#">SCT-Distil-model-wangchanberta</a>        | 56.23 | 65.18  |             |
| <a href="#">SCT-Distil-model-phayathaibert</a>        | 58.32 | 67.42  |             |
| <a href="#">ConGen-model-XLMR</a>                     | 60.29 | 68.56  |             |
| <a href="#">ConGen-model-wangchanberta</a>            | 59.11 | 67.42  |             |
| <a href="#">ConGen-model-phayathaibert</a>            | 59.24 | 67.69  |             |
| <a href="#">ConGen-BGE_M3-model-phayathaibert</a>     | 83.36 | 88.29  |             |
| <a href="#">distiluse-base-multilingual-cased-v2</a>  | 32.50 | 42.20  | ✓           |
| <a href="#">paraphrase-multilingual-mpnet-base-v2</a> | 54.39 | 63.12  | ✓           |
| <a href="#">BGE M-3</a>                               | 89.12 | 93.43  | ✓           |
| <a href="#">Cohere-embed-multilingual-v2.0</a>        | 85.45 | 90.33  | ✓           |



Sentence Similarity

sentence-transformers



PyTorch



ONNX

xlm-roberta



License: mit



Model card



Files



Community 47

[Edit model card](#)

For more details please refer to our github repo:

<https://github.com/FlagOpen/FlagEmbedding>

## BGE-M3 ([paper](#), [code](#))

In this project, we introduce BGE-M3, which is distinguished for its versatility in Multi-Functionality, Multi-Linguality, and Multi-Granularity.

- Multi-Functionality: It can simultaneously perform the three common retrieval functionalities of embedding model: dense retrieval, multi-vector retrieval, and sparse retrieval.
- Multi-Linguality: It can support more than 100 working languages.
- Multi-Granularity: It is able to process inputs of different granularities, spanning from short sentences to long documents of up to 8192 tokens.



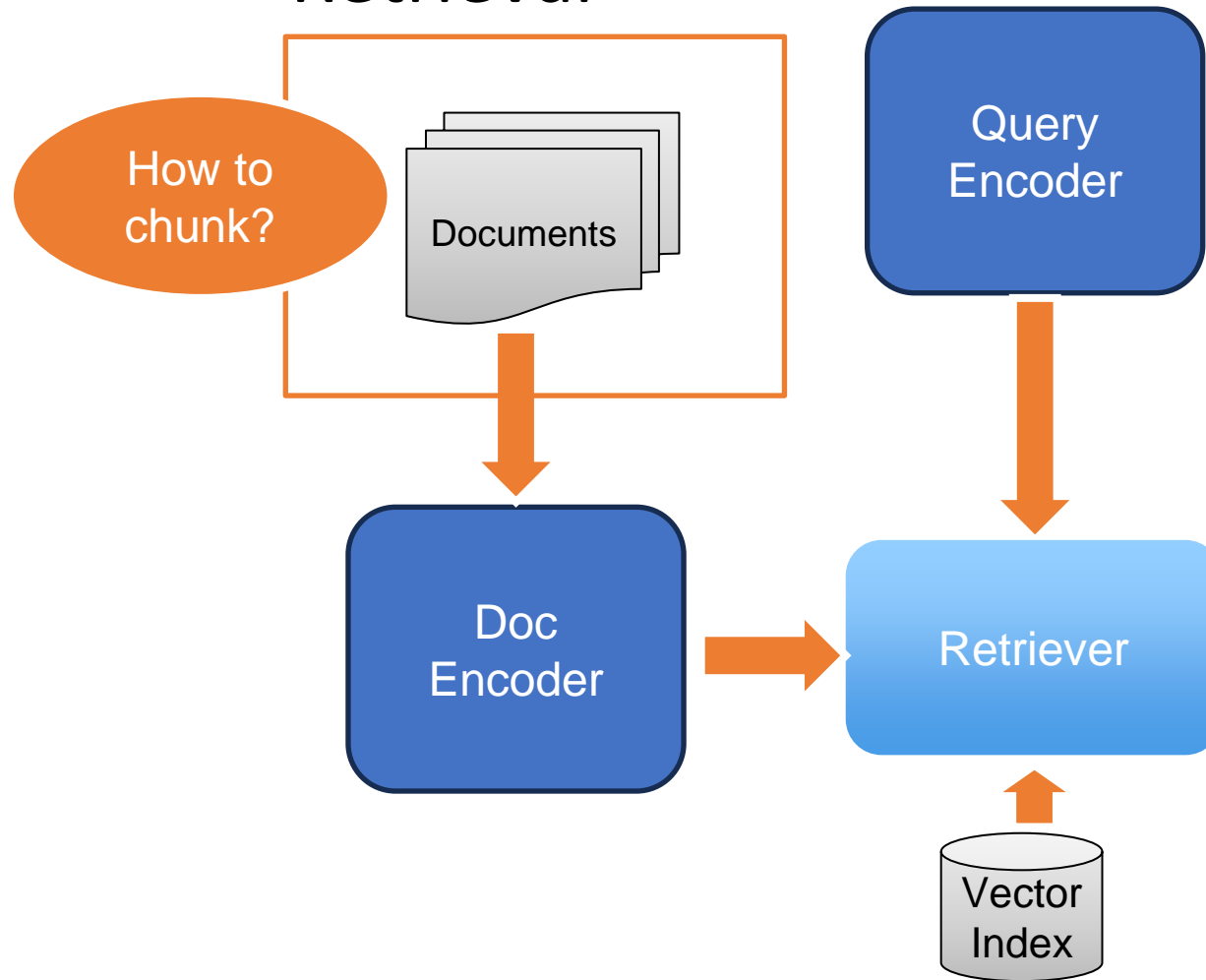
• Model

| Model Name                               | Sequence  |        | Introduction   |
|--|-----------|--------|--|
|  | Dimension | Length |  |
| <a href="#">BAAI/bge-m3</a>              | 1024      | 8192   | multilingual; unified fine-tuning (dense, sparse, and colbert) from bge-m3-unsupervised  |
| <a href="#">BAAI/bge-m3-unsupervised</a> | 1024      | 8192   | multilingual; contrastive learning from bge-m3-retromae  |
| <a href="#">BAAI/bge-m3-retromae</a>     | --        | 8192   | multilingual; extend the max_length of <a href="#">xlm-roberta</a> to 8192 and further pretrained via <a href="#">retromae</a> |
| <a href="#">BAAI/bge-large-en-v1.5</a>   | 1024      | 512    | English model  |
| <a href="#">BAAI/bge-base-en-v1.5</a>    | 768       | 512    | English model  |
| <a href="#">BAAI/bge-small-en-v1.5</a>   | 384       | 512    | English model  |

# Huggingface download model

1. `pip install hf_transfer`
2. `export HF_HUB_ENABLE_HF_TRANSFER=1`
3. `huggingface-cli download <model-name> --local-dir <model-dir>`

# Retrieval



## จังหวัดพระนครศรีอยุธยา

"พระนครศรีอยุธยา" เปลี่ยนนามาที่นี่ สำหรับความหมายอื่น ดูที่ [อยุธยา](#)

**พระนครศรีอยุธยา** เป็นจังหวัดหนึ่งในภาคกลางและเป็นเขตเศรษฐกิจอุตสาหกรรมที่สำคัญ โดยมีผลิตภัณฑ์มวลรวมของจังหวัดมีมูลค่าสูงเป็นอันดับ 3 ของประเทศ<sup>[3]</sup> และมีประวัติศาสตร์ความเป็นมาที่ยาวนาน เคยมีชื่อเสียงเป็นแหล่งปลูกข้าวที่สำคัญ จังหวัดพระนครศรีอยุธยาเป็นจังหวัดเดียวในประเทศไทยที่ไม่มีอำเภอเมือง แต่มี [อำเภอพระนครศรีอยุธยา](#) เป็นศูนย์กลางการบริหารจัดการด้านต่าง ๆ ชาวบ้านโดยทั่วไปนิยมเรียกอีกชื่อว่า "กรุงเก่า" หรือ "เมืองกรุงเก่า" ห่างจากกรุงเทพมหานครประมาณ 75 กิโลเมตร

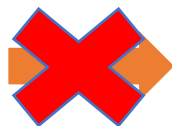
### ประวัติศาสตร์[แก้]

บทความหลัก: [อาณาจักรอยุธยา](#)

พระนครศรีอยุธยาเคยเป็น**ราชธานี** (เมืองหลวง) ของ**อาณาจักรอยุธยา** หรืออาณาจักรสยาม ตลอดระยะเวลา 417 ปี ตั้งแต่ พ.ศ. 1893 กระทั่งเสียกรุงแก่พม่าเมื่อ พ.ศ. 2310 ครั้นเมื่อ**พระเจ้าตากสินมหาราช**ทรงสถาปนาราชธานีแห่งใหม่ที่**กรุงธนบุรี** กรุงศรีอยุธยาก็ไม่ได้กลายเป็นเมืองร้าง ยังมีคนที่รักถิ่นฐานบ้านเดิมอาศัยอยู่และมีราษฎรที่หลบหนีไปอยู่ตามปากลับเข้ามาอาศัยอยู่รอบ ๆ เมือง รวมกันเข้าเป็นเมือง จนทางการยกเป็นเมืองจัตวาเรียกว่า "เมืองกรุงเก่า"

เมื่อ พ.ศ. 2325 ในรัชสมัย**พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช** ทรงยกเมืองกรุงเก่าขึ้นเป็น หัวเมืองจัตวา เช่นเดียวกับในสมัยกรุงธนบุรี หลังจากนั้น**พระบาทสมเด็จพระจุลจอมเกล้าเจ้าอยู่หัว**โปรดให้จัดการปฏิรูปการปกครองทั้งส่วนกลางและส่วนภูมิภาค โดยการปกครองส่วนภูมิภาคนั้น โปรดให้จัดการปกครองแบบเทศาภิบาลขึ้น โดยให้รวมเมืองที่อยู่ใกล้เคียงกัน 3-4 เมือง ขึ้นเป็น**มณฑล** มีข้าหลวงเทศาภิบาลเป็นผู้ปกครอง โดยในปี พ.ศ. 2438 โปรดให้จัดตั้งมณฑลกรุงเก่าขึ้น ประกอบด้วยหัวเมืองต่าง ๆ คือ กรุงเก่าหรือ**อยุธยา** **อ่างทอง** **สระบุรี** **พระพุทธรักษา** **ลพบุรี** **พรหมบุรี** **อินทร์บุรี** และ**สิงห์บุรี**

ต่อมาทรงพระกรุณาโปรดเกล้าฯ ให้รวมเมืองอินทร์บุรีและเมืองพรหมบุรีเข้ากับเมืองสิงห์บุรี รวมเมืองพระพุทธรักษาเข้ากับเมืองสระบุรี ตั้งที่ว่าการมณฑลที่อยุธยา และในปี พ.ศ. 2469 เปลี่ยนชื่อจาก "มณฑลกรุงเก่า" เป็น "มณฑลอยุธยา" ซึ่งจากการจัดตั้งมณฑลอยุธยามีผลให้อยุธยามีความสำคัญทางการบริหารการปกครองมากขึ้น การสร้างสิ่งสาธารณูปโภคหลายอย่างมีผลต่อการพัฒนาเมืองอยุธยาในเวลาต่อมา จนเมื่อยกเลิกการปกครองระบบ**มณฑลเทศาภิบาล** ภายหลัง**การเปลี่ยนแปลงการปกครอง พ.ศ. 2475** อยุธยาจึงเปลี่ยนฐานะเป็นจังหวัด





## จังหวัดพระนครศรีอยุธยา

"พระนครศรีอยุธยา" เปลี่ยนนามาที่นี่ สำหรับความหมายอื่น ดูที่ [อยุธยา](#)

**พระนครศรีอยุธยา** เป็นจังหวัดหนึ่งในภาคกลางและเป็นเขตเศรษฐกิจอุตสาหกรรมที่สำคัญ โดยมีผลิตภัณฑ์มวลรวมของจังหวัดมีมูลค่าสูงเป็นอันดับ 3 ของประเทศ<sup>[3]</sup> และมีประวัติศาสตร์ความเป็นมาที่ยาวนาน เคยมีชื่อเสียงเป็นแหล่งปลูกข้าวที่สำคัญ จังหวัดพระนครศรีอยุธยาเป็นจังหวัดเดียวในประเทศไทยที่ไม่มีอำเภอเมือง แต่มี [อำเภอพระนครศรีอยุธยา](#) เป็นศูนย์กลางการบริหารจัดการด้านต่าง ๆ ชาวบ้านโดยทั่วไปนิยมเรียกอีกชื่อว่า "กรุงเก่า" หรือ "เมืองกรุงเก่า" ห่างจากกรุงเทพมหานครประมาณ 75 กิโลเมตร

### ประวัติศาสตร์[แก้]

บทความหลัก: [อาณาจักรอยุธยา](#)

พระนครศรีอยุธยาเคยเป็น**ราชธานี** (เมืองหลวง) ของ**อาณาจักรอยุธยา** หรืออาณาจักรสยาม ตลอดระยะเวลา 417 ปี ตั้งแต่ พ.ศ. 1893 กระทั่งเสียกรุงแก่พม่าเมื่อ พ.ศ. 2310 ครั้นเมื่อ**พระเจ้าตากสินมหาราช**ทรงสถาปนาราชธานีแห่งใหม่ที่**กรุงธนบุรี** กรุงศรีอยุธยาที่ไม่ได้กลายเป็นเมืองร้าง ยังมีคนที่รักถิ่นฐานบ้านเดิมอาศัยอยู่และมีราษฎรที่หลบหนีไปอยู่ตามปากลับเข้ามาอาศัยอยู่รอบ ๆ เมือง รวมกันเข้าเป็นเมือง จนทางการยกเป็นเมืองจัดว่าเรียกว่า "เมืองกรุงเก่า"

เมื่อ พ.ศ. 2325 ในรัชสมัย**พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช** ทรงยกเมืองกรุงเก่าขึ้นเป็น หัวเมืองจัตวา เช่นเดียวกับในสมัยกรุงธนบุรี หลังจากนั้น**พระบาทสมเด็จพระจุลจอมเกล้าเจ้าอยู่หัว**โปรดให้จัดการปฏิรูปการปกครองทั้งส่วนกลางและส่วนภูมิภาค โดยการปกครองส่วนภูมิกาคนั้น โปรดให้จัดการปกครองแบบเทศาภิบาลขึ้น โดยให้รวมเมืองที่อยู่ใกล้เคียงกัน 3-4 เมือง ขึ้นเป็น**มณฑล** มีข้าหลวงเทศาภิบาลเป็นผู้ปกครอง โดยในปี พ.ศ. 2438 โปรดให้จัดตั้งมณฑลกรุงเก่าขึ้น ประกอบด้วยหัวเมืองต่าง ๆ คือ กรุงเก่าหรือ**อยุธยา** **อ่างทอง** **สระบุรี** **พระพุทธรบาท** **ลพบุรี** **พรหมบุรี** **อินทร์บุรี** และ**สิงห์บุรี**

ต่อมาทรงพระกรุณาโปรดเกล้าฯ ให้รวมเมืองอินทร์บุรีและเมืองพรหมบุรีเข้ากับเมืองสิงห์บุรี รวมเมืองพระพุทธรบาทเข้ากับเมืองสระบุรี ตั้งที่ว่ากรมณฑลที่อยุธยา และในปี พ.ศ. 2469 เปลี่ยนชื่อจาก "มณฑลกรุงเก่า" เป็น "มณฑลอยุธยา" ซึ่งจากการจัดตั้งมณฑลอยุธยามีผลให้อยุธยามีความสำคัญทางการบริหารการปกครองมากขึ้น การสร้างสิ่งสาธารณูปโภคหลายอย่างมีผลต่อการพัฒนาเมืองอยุธยาในเวลาต่อมา จนเมื่อยกเลิกการปกครองระบบ**มณฑลเทศาภิบาล** ภายหลัง**การเปลี่ยนแปลงการปกครอง พ.ศ. 2475** อยุธยาจึงเปลี่ยนฐานะเป็นจังหวัด

BAAI/bge-m3

[0.016539733856916428,  
0.03937997296452522,  
-0.010712525807321072,  
-0.02129676751792431,  
-0.025445323437452316]

## จังหวัดพระนครศรีอยุธยา

"พระนครศรีอยุธยา" เปลี่ยนทางมาที่นี่ สำหรับความหมายอื่น ดูที่ [อยุธยา](#)

พระนครศรีอยุธยา เป็นจังหวัดหนึ่งในภาคกลางและเป็นเขตเศรษฐกิจอุตสาหกรรมที่สำคัญ โดยมีผลิตภัณฑ์มวลรวมของจังหวัดมีมูลค่าสูงเป็นอันดับ 3 ของประเทศ<sup>[3]</sup> และมีประวัติศาสตร์ความเป็นมาที่ยาวนาน เคยมีชื่อเสียงเป็นแหล่งปลูกข้าวที่สำคัญ จังหวัดพระนครศรีอยุธยาเป็นจังหวัดเดียวในประเทศไทยที่ไม่มีอำเภอเมือง แต่มี [อำเภอพระนครศรีอยุธยา](#) เป็นศูนย์กลางการบริหารจัดการด้านต่าง ๆ ชาวบ้านโดยทั่วไปนิยมเรียกอีกชื่อว่า "กรุงเก่า" หรือ "เมืองกรุงเก่า" ห่างจากกรุงเทพมหานครประมาณ 75 กิโลเมตร

## ประวัติศาสตร์[แก้]

บทความหลัก: [อาณาจักรอยุธยา](#)

พระนครศรีอยุธยาเคยเป็นราชธานี (เมืองหลวง) ของอาณาจักรอยุธยา หรือ อาณาจักรสยาม ตลอดระยะเวลา 417 ปี ตั้งแต่ พ.ศ. 1893 กระทั่งเสียกรุงแก่พม่าเมื่อ พ.ศ. 2310 ครั้นเมื่อพระเจ้าตากสินมหาราชทรงสถาปนาราชธานีแห่งใหม่ที่กรุงธนบุรี กรุงศรีอยุธยาก็ไม่ได้กลายเป็นเมืองร้าง ยังมีคนที่ยึดมั่นฐานบ้านเดิมอาศัยอยู่และมีราษฎรที่หลบหนีไปอยู่ตามป่ากลับเข้ามาอาศัยอยู่รอบ ๆ เมือง รวมกันเข้าเป็นเมือง จนทางการยกเป็นเมืองจัตวาเรียกว่า "เมืองกรุงเก่า"

เมื่อ พ.ศ. 2325 ในรัชสมัยพระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช ทรงยกเมืองกรุงเก่าขึ้นเป็น หัวเมืองจัตวา เช่นเดียวกับในสมัยกรุงธนบุรี หลังจากนั้นพระบาทสมเด็จพระจุลจอมเกล้าเจ้าอยู่หัวโปรดให้จัดการปฏิรูปการปกครองทั้งส่วนกลางและส่วนภูมิภาค โดยการปกครองส่วนภูมิกษัตริย์นั้น โปรดให้จัดการปกครองแบบเทศาภิบาลขึ้น โดยให้รวมเมืองที่อยู่ใกล้เคียงกัน 3-4 เมือง ขึ้นเป็นมณฑล มีข้าหลวงเทศาภิบาลเป็นผู้ปกครอง โดยในปี พ.ศ. 2438 โปรดให้จัดตั้งมณฑลกรุงเก่าขึ้น ประกอบด้วยหัวเมืองต่าง ๆ คือ กรุงเก่าหรืออยุธยา อ่างทอง สระบุรี พระพทุธนาท ลพบุรี พรหมบุรี อินทร์บุรี และสิงห์บุรี

ต่อมาทรงพระกรุณาโปรดเกล้าฯ ให้รวมเมืองอินทร์บุรีและเมืองพรหมบุรีเข้ากับเมืองสิงห์บุรี รวมเมืองพระพทุธนาทเข้ากับเมืองสระบุรี ตั้งที่ว่าการมณฑลที่อยุธยา และในปี พ.ศ. 2469 เปลี่ยนชื่อจาก "มณฑลกรุงเก่า" เป็น "มณฑลอยุธยา" ซึ่งจากการจัดตั้งมณฑลอยุธยามีผลให้อยุธยามีความสำคัญทางด้านการบริหารการปกครองมากขึ้น การสร้างสิ่งสาธารณูปโภคหลายอย่างมีผลต่อการพัฒนาเมืองอยุธยาในเวลาต่อมา จนเมื่อยกเลิกการปกครองระบบมณฑลเทศาภิบาล ภายหลังการเปลี่ยนแปลงการปกครอง พ.ศ. 2475 อยุธยาจึงเปลี่ยนฐานะเป็นจังหวัด

Encoder

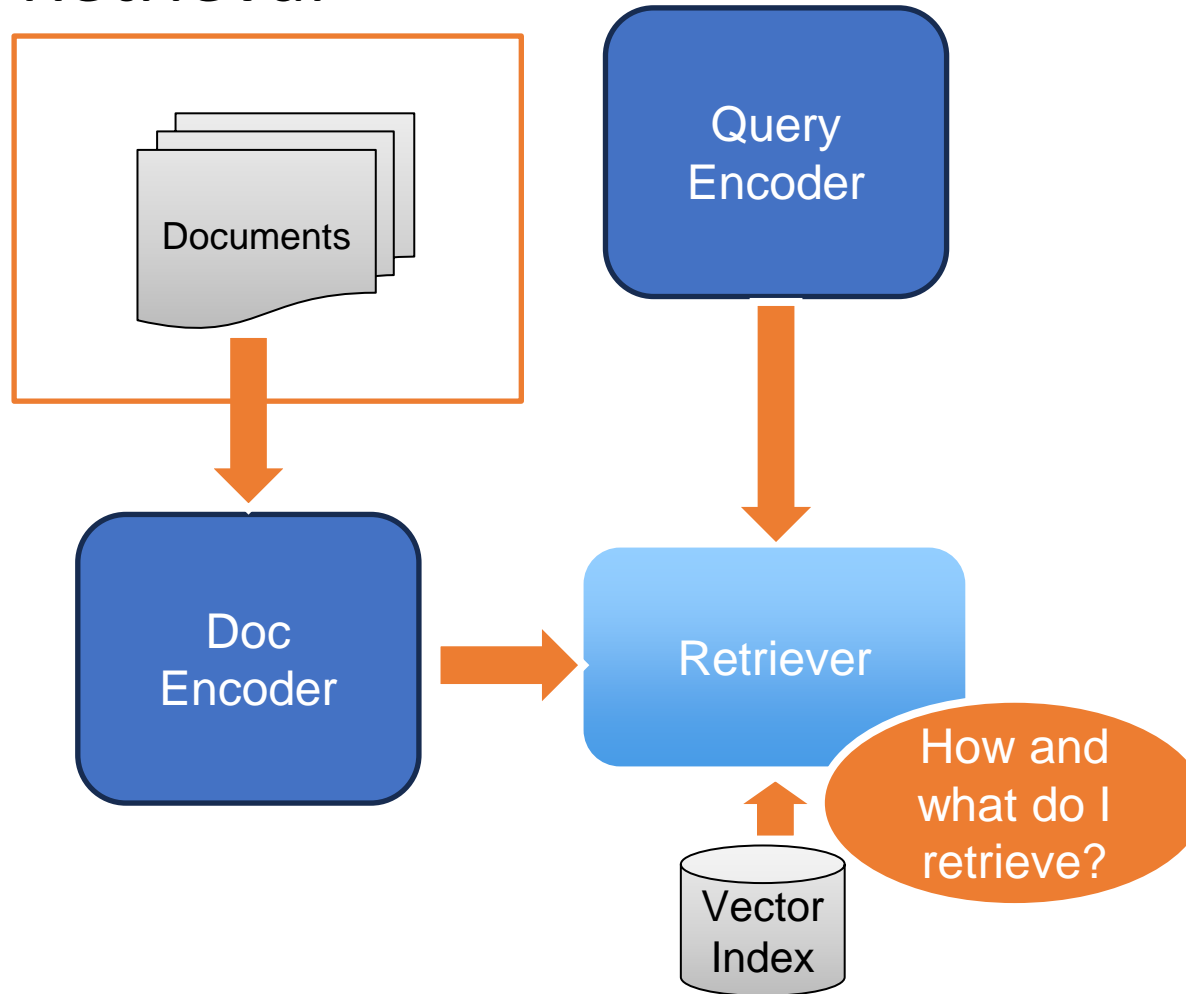
[0.005539661273360252,  
0.015187489800155163,  
-0.01164360623806715,  
-0.01728619821369648,  
-0.021985329687595367]

[0.01168364193290472,  
0.04119976609945297,  
-0.00417240709066391,  
-0.02726016566157341,  
-0.03603764623403549]

[0.031089192256331444,  
0.04272811487317085,  
-0.00500113982707262,  
0.0019403527257964015,  
-0.01635820046067238]

[0.021662477403879166,  
0.04273651912808418,  
0.01920708455145359,  
0.007724749855697155, -  
0.007433386519551277]

# Retrieval



# Vector Database

[https://docs.llamaindex.ai/en/stable/module\\_guides/storing/vector\\_stores/#vector-store-options--feature-support](https://docs.llamaindex.ai/en/stable/module_guides/storing/vector_stores/#vector-store-options--feature-support)

| Vector Store              | Type                | Metadata Filtering | Hybrid Search | Delete | Store Documents | Async |
|---------------------------|---------------------|--------------------|---------------|--------|-----------------|-------|
| Apache Cassandra®         | self-hosted / cloud | ✓                  |               | ✓      | ✓               |       |
| Astra DB                  | cloud               | ✓                  |               | ✓      | ✓               |       |
| Azure AI Search           | cloud               | ✓                  | ✓             | ✓      | ✓               |       |
| Azure CosmosDB<br>MongoDB | cloud               |                    |               | ✓      | ✓               |       |
| BaiduVectorDB             | cloud               | ✓                  | ✓             |        | ✓               |       |
| ChatGPT Retrieval Plugin  | aggregator          |                    |               | ✓      | ✓               |       |
| Chroma                    | self-hosted         | ✓                  |               | ✓      | ✓               |       |
| Couchbase                 | self-hosted / cloud | ✓                  | ✓             | ✓      | ✓               |       |
| DashVector                | cloud               | ✓                  | ✓             | ✓      | ✓               |       |
| Databricks                | cloud               | ✓                  |               | ✓      | ✓               |       |
| Deeplake                  | self-hosted / cloud | ✓                  |               | ✓      | ✓               |       |
| FAISS                     | in-memory           |                    |               |        |                 |       |

# FAISS

Based on “semantic similarity” (i.e., cosine score between embeddings)

## **Billion-scale similarity search with GPUs**

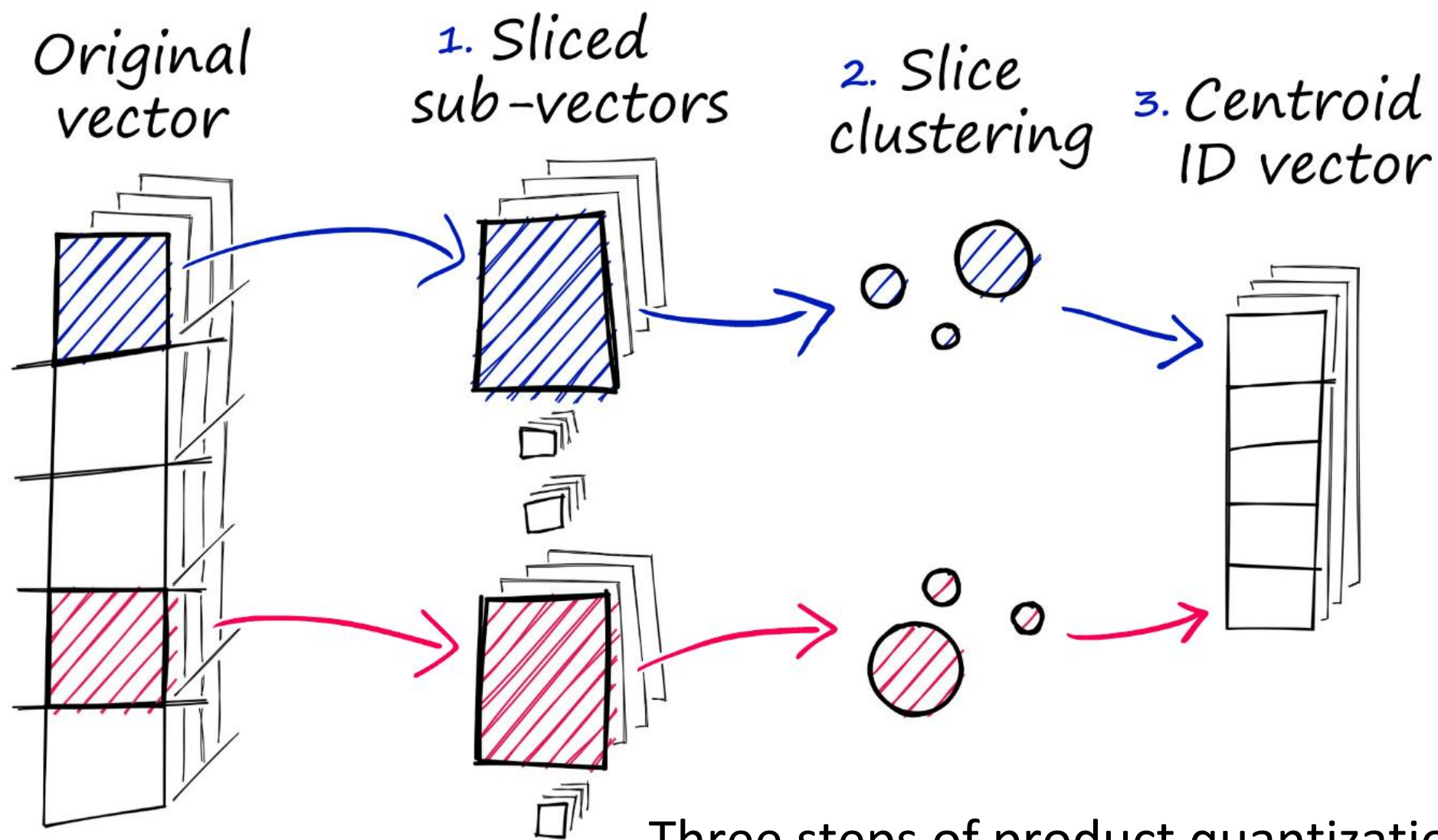
Jeff Johnson  
Facebook AI Research  
New York

Matthijs Douze  
Facebook AI Research  
Paris

Hervé Jégou  
Facebook AI Research  
Paris

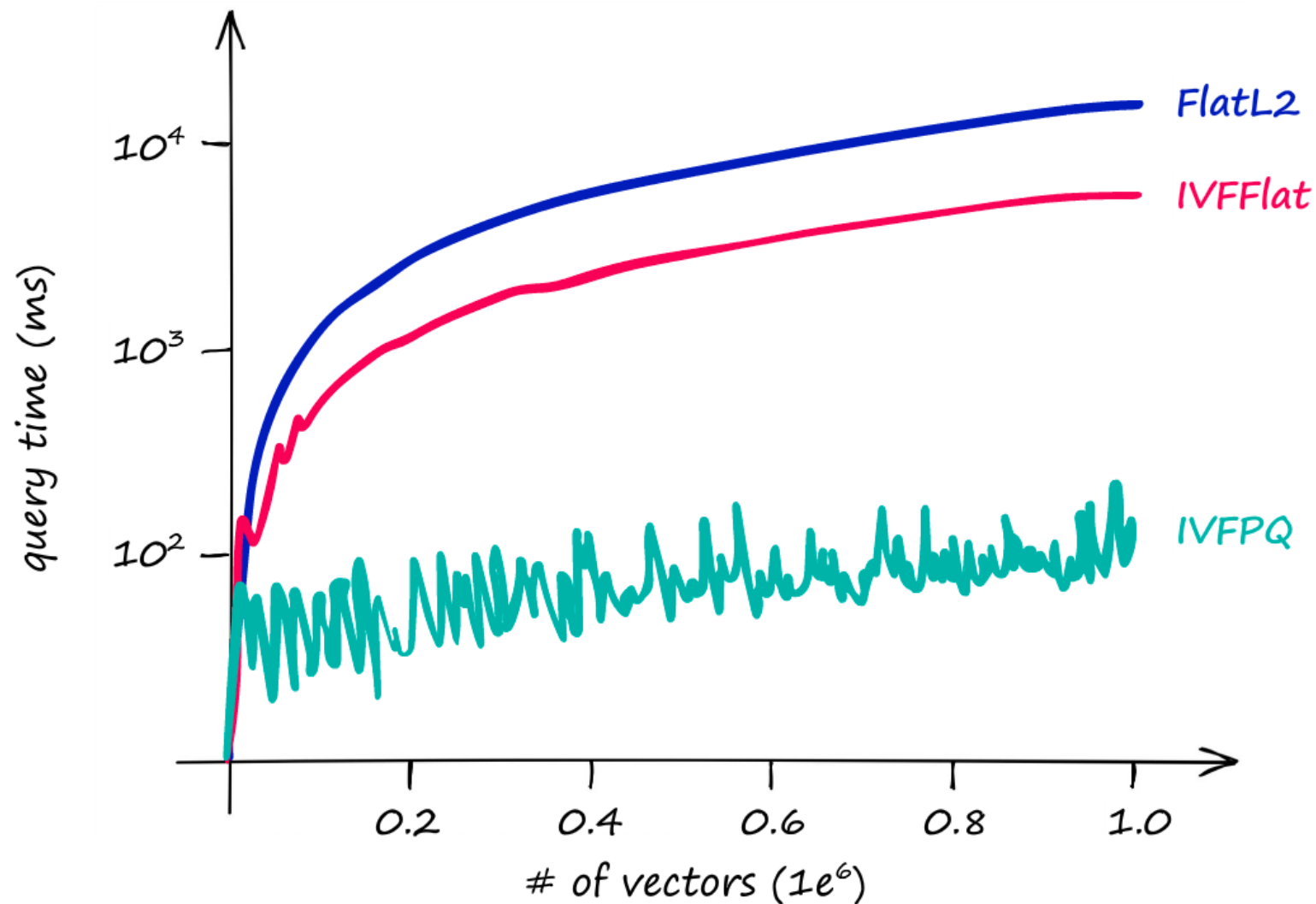
<https://github.com/facebookresearch/faiss>

# FAISS



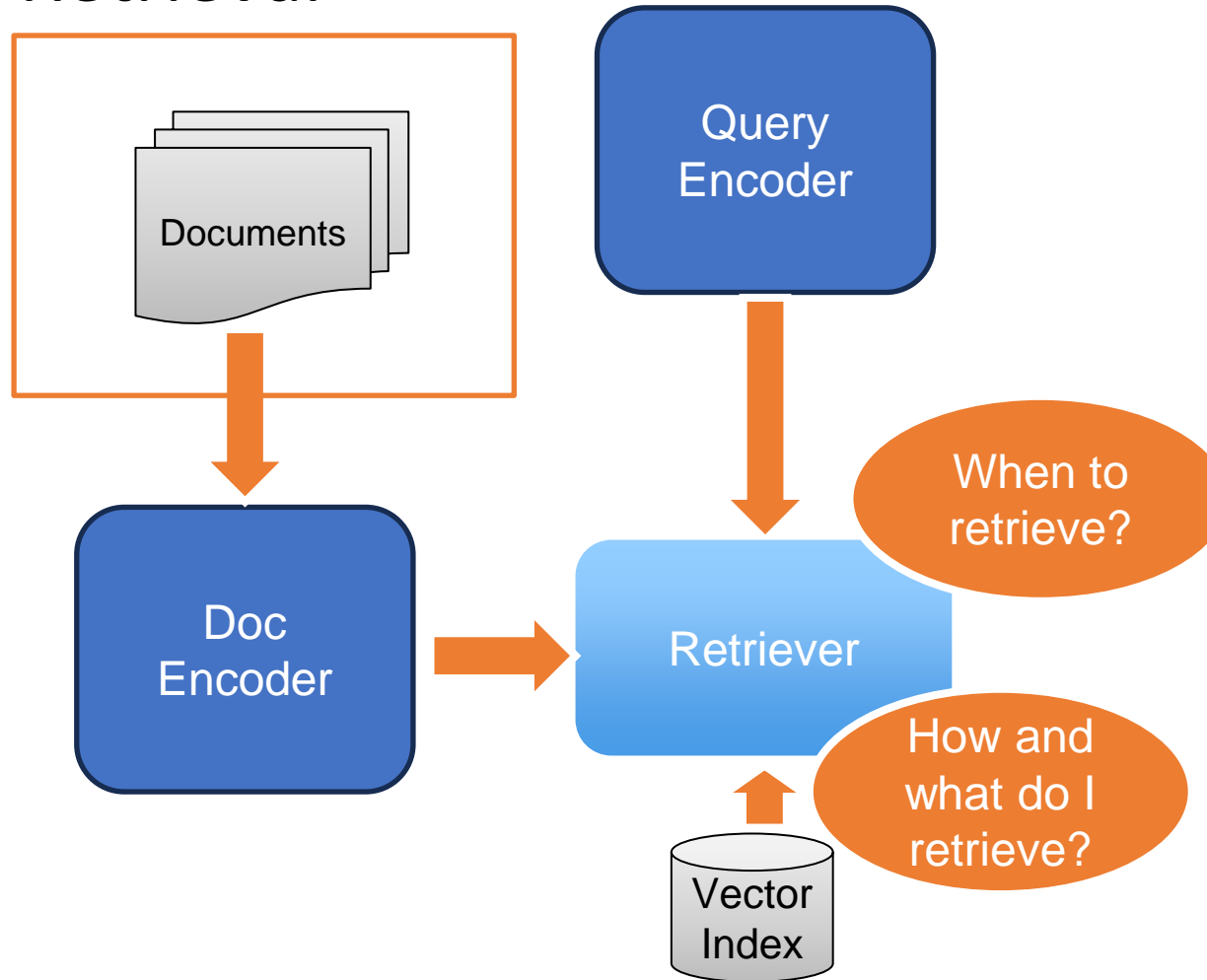
Three steps of product quantization

# FAISS



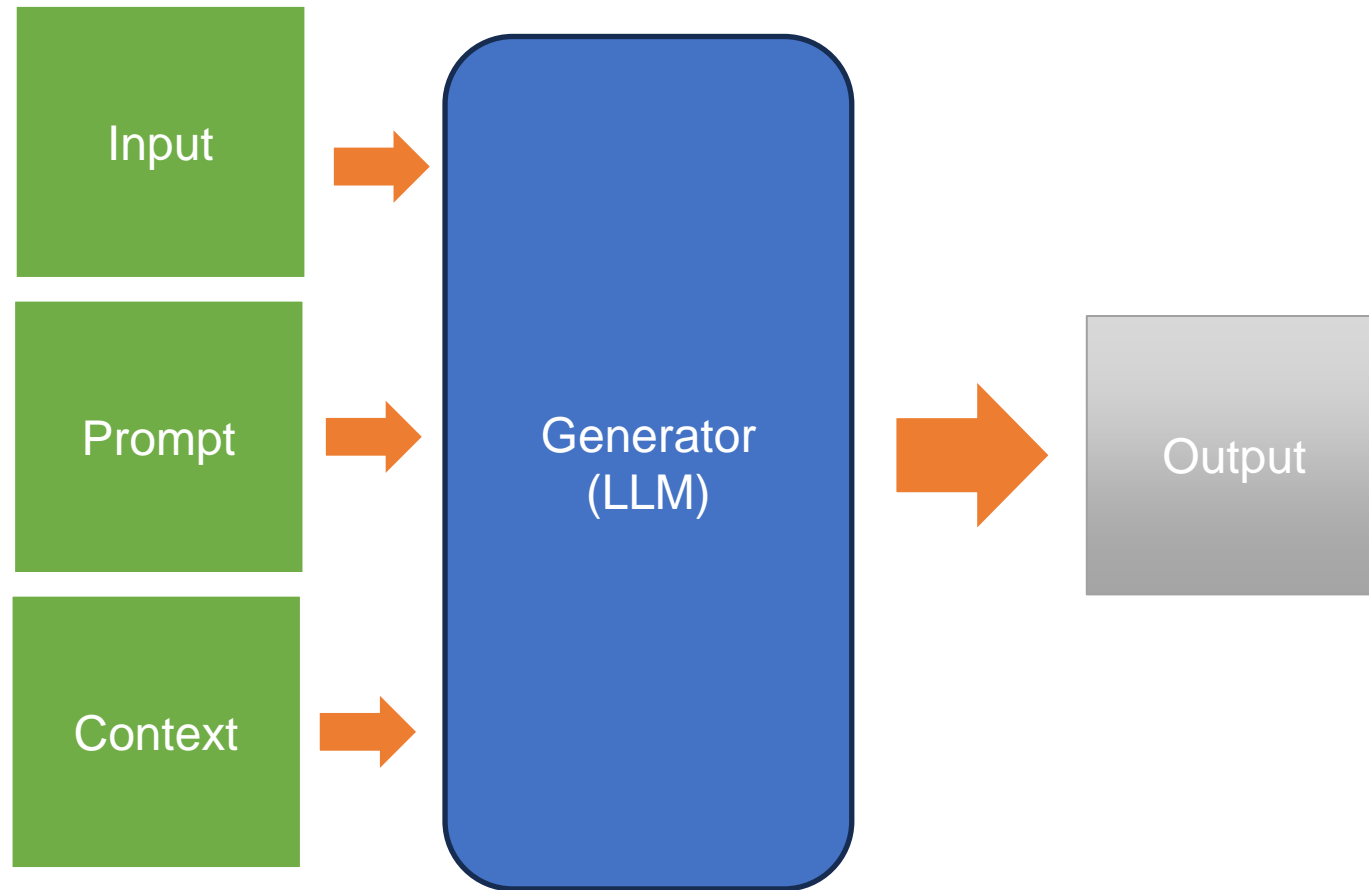
Query time / number of vectors for the three indexes

# Retrieval





# Generation (LLM)



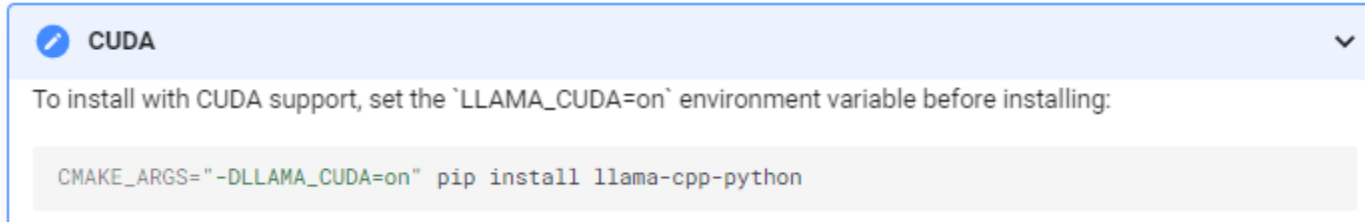
# llama.cpp

- <https://github.com/ggerganov/llama.cpp>
  - Inference of Meta's [LLaMA](#) model (and others) in pure C/C++
- <https://github.com/abetlen/llama-cpp-python>
  - Python Bindings for llama.cpp

# Install llama.cpp

<https://llama-cpp-python.readthedocs.io/en/stable/>

- Use CuBLAS if you have CUDA and an NVidia GPU

A screenshot of a documentation box with a light blue header. The header contains a blue checkmark icon and the word "CUDA". Below the header, the text reads: "To install with CUDA support, set the `LLAMA\_CUDA=on` environment variable before installing:". Below this text is a light gray code block containing the command: 

```
CMAKE_ARGS="-DLLAMA_CUDA=on" pip install llama-cpp-python
```

**CUDA**

To install with CUDA support, set the `LLAMA\_CUDA=on` environment variable before installing:

```
CMAKE_ARGS="-DLLAMA_CUDA=on" pip install llama-cpp-python
```

- Use METAL if you are running on an M1/M2 MacBook
- Use CLBlast if you are running on an AMD/Intel GPU

# Install llama.cpp

<https://llama-cpp-python.readthedocs.io/en/stable/>

- Use CuBLAS if you have CUDA and an NVidia GPU
- Use METAL if you are running on an M1/M2 MacBook

 Metal 

To install with Metal (MPS), set the `LLAMA\_METAL=on` environment variable before installing:

```
CMAKE_ARGS="-DLLAMA_METAL=on" pip install llama-cpp-python
```

**\*\*Pre-built Wheel (New)\*\*** It is also possible to install a pre-built wheel with Metal support. As long as your system meets some requirements: - MacOS Version is 11.0 or later - Python Version is 3.10, 3.11 or 3.12

```
pip install llama-cpp-python \
  --extra-index-url https://abetlen.github.io/llama-cpp-python/whl/metal
```

- Use CLBlast if you are running on an AMD/Intel GPU

# llama.cpp

(<https://github.com/ggerganov/llama.cpp>)

```
git clone https://github.com/ggerganov/llama.cpp.git  
cd llama.cpp  
make
```

Try:

```
./main -m <path to model>/openthaigpt-Q4_K_M.gguf -p "Building a website  
can be done in 10 simple steps:\nStep 1:" -n 400 -e
```

```
./main -m <path to model>/openthaigpt-Q4_K_M.gguf -p "สูตรทำส้มตำปูปลาร้า  
ใน 10 ขั้นตอน\nขั้นตอนที่ 1:" -n 400 -e
```

# llama.cpp

`./server -m <path to model>/openthetaigt-Q4_K_M.gguf`

Go to: 127.0.0.1:8080

**llama.cpp**

☒ Chat ☐ Completion

Prompt

This is a conversation between User and Llama, a friendly chatbot. Llama is helpful, kind, honest, good at writing, and never fails to answer any requests immediately and with precision.

User name  Bot name

Prompt template

{{prompt}}  
{{history}}  
{{char}}:

Chat history template

{{name}}: {{message}}

Grammar

Use gbnf or JSON Schema+convert

order: prop1,prop2,prop3

Predictions  400 Temperature  0.7

Penalize repeat sequence  1.18 Consider N tokens for penalize  256

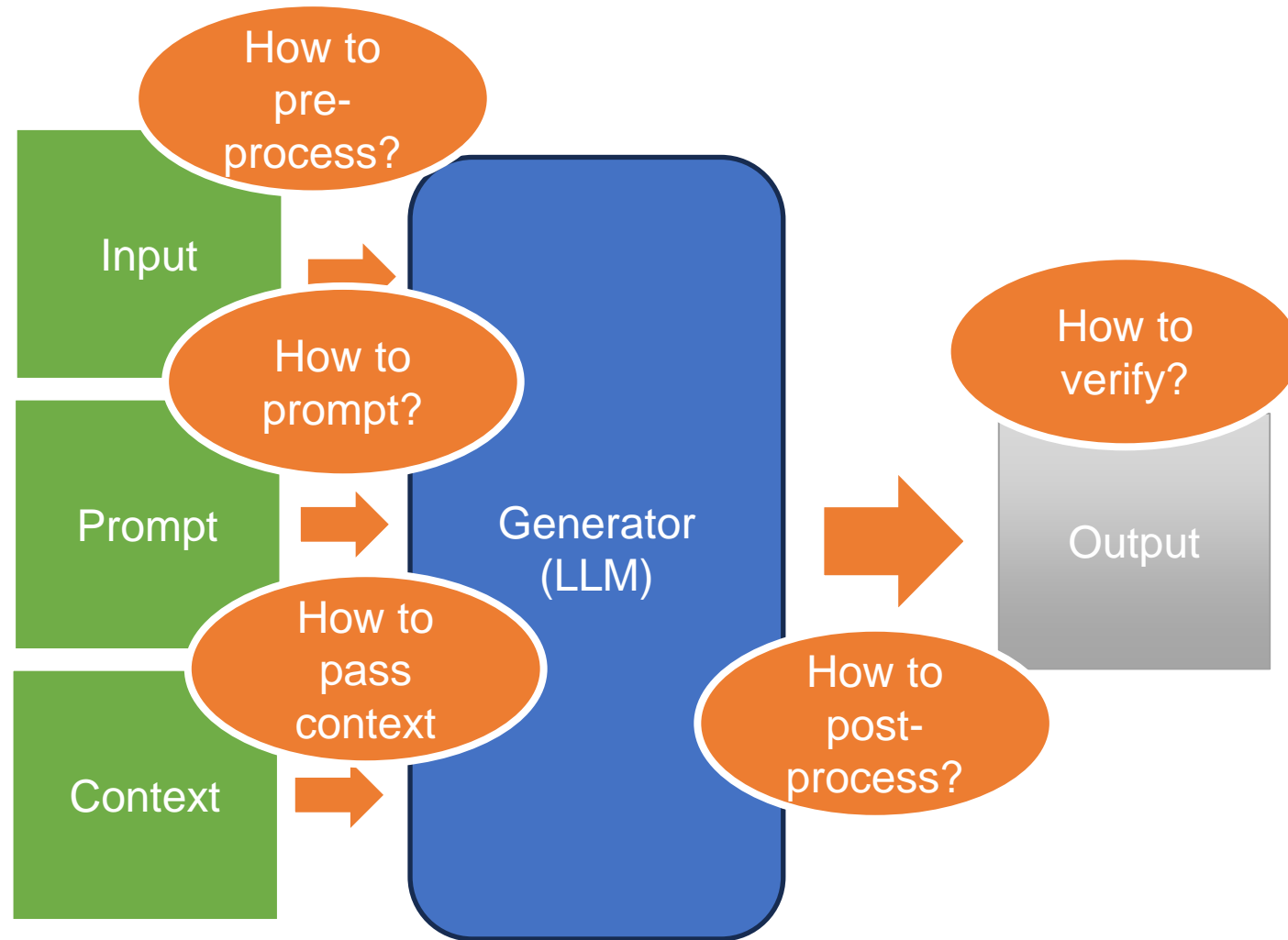
Penalize repetition of newlines ☐ Top-K sampling  40

Top-P sampling  0.95 Min-P sampling  0.05

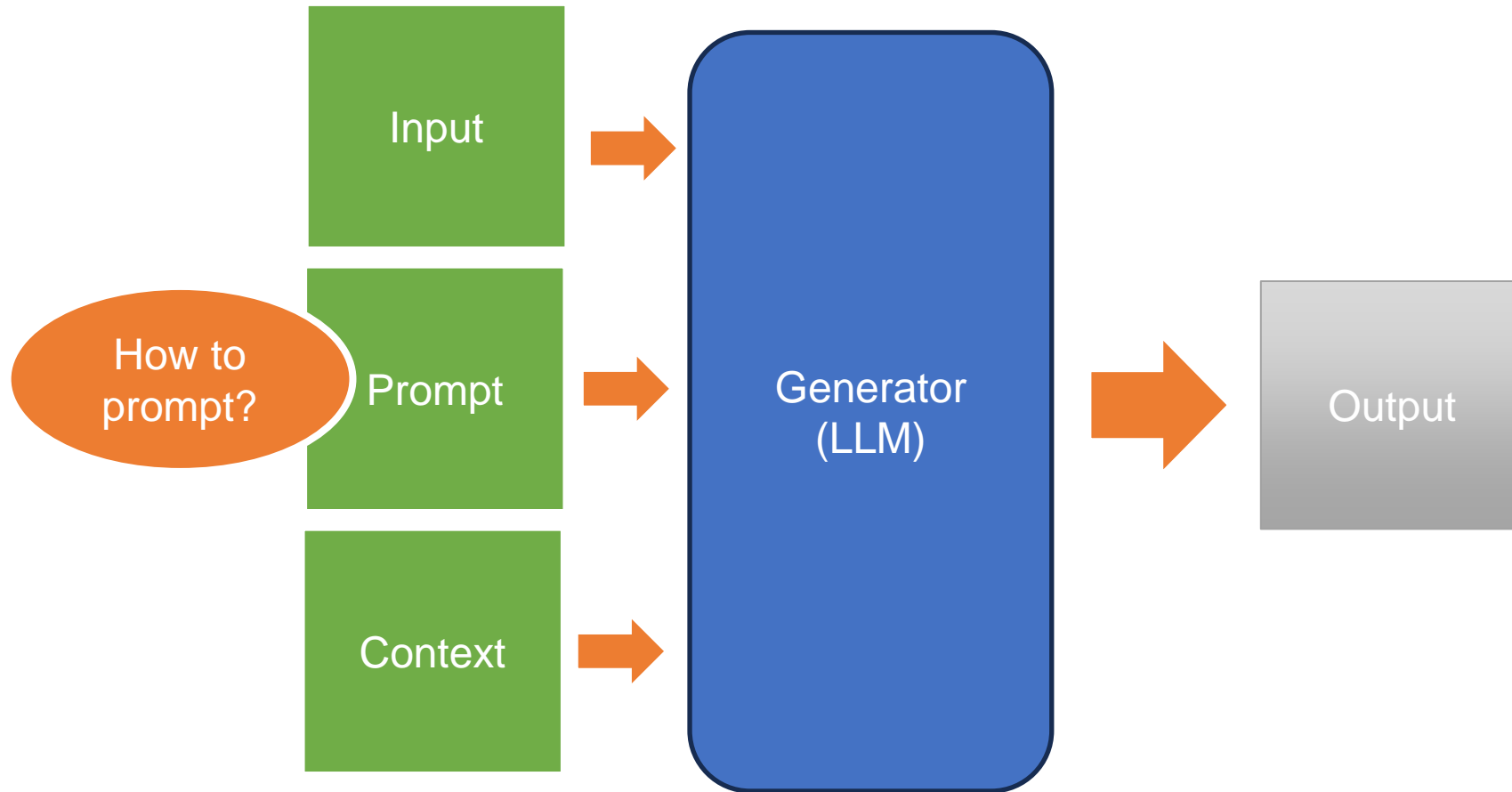
Say something...

Powered by [llama.cpp](#) and [ggml.ai](#)

# Generation (LLM)



# Generation (LLM)





# How to prompt?

## Prompt

This is a conversation between User and Llama, a friendly chatbot. Llama is helpful, kind, honest, good at writing, and never fails to answer any requests immediately and with precision.

## OpenThaiGPT

You are a question answering assistant. Answer the question as truthful and helpful as possible คุณคือผู้ช่วยตอบคำถาม จงตอบคำถามอย่างถูกต้องและมีประโยชน์ที่สุด

# Chat

User: สวัสดี

OPT: สวัสดีจ้า มีเรื่องอะไรให้ช่วยมั๊ยคะ

User: ทำอะไรได้มั่ง

OPT: ทำได้เยอะเลยคะ ตั้งแต่เล่นเกม ดูหนัง อ่านหนังสือ แต่งกลอน คิดเลข คำนวณ [/INST]

User: บอกวิธีทำสัมผัสน้อย

OPT: ได้ค่ะ!

User: ว่ามา

OPT: ใส่ปุ้ มะเขือเทศ พริกหยวก แครอท ผักชี กุ้งแห้งและawang

User: ไม่เอาแครอทได้มะ

OPT: ได้ค่ะ หรือจะเอาเป็นกะหล่ำปลีแทนก็ได้้นะ

User: ไม่เอากระหล่ำปลีด้วย

OPT: ได้ค่ะ มาปรุงสัมผัสน้อยเลย

User: มาๆ

OPT: เริ่มเลย! 1. ตักน้ำตาลทรายใส่แก้ว

2. เติมน้ำเปล่าลงไป

3. คนให้น้ำตาลทรายละลาย

4. เทน้ำลงในกะหล่ำปลีที่ต้มไว้

5. ใส่เนื้อสัตว์ต่างๆ ลงไป

6. ผัดให้เข้ากัน

7. ตักขึ้นราดบนข้าวเหนียว

User: แน่ใจว่านี่สัมผั


OPT: มันก็เหมือนสัมผัสน้อยนา

# Completion

Reset all to default ☐ Chat ☒ Completion

Prompt

สูตรทำส้มตำปูปลาร้าใน 10 ขั้นตอน  
ขั้นตอนที่ 1:





สูตรทำส้มตำปูปลาร้าใน 10 ขั้นตอน  
ขั้นตอนที่ 1: การเตรียมวัตถุดิบ ขั้นตอนที่ 2: สับสับ ขั้นตอนที่ 3: การตั้งค่าน้ำจิ้ม ขั้นตอนที่ 4: การ  
ทำความสะอาด ขั้นตอนที่ 5: การหั่นมะเขือเทศ ขั้นตอนที่ 6: การสับปลาร้า ขั้นตอนที่ 7: การหั่นพริก  
ขั้นตอนที่ 8: การผัดส่วนผสม ขั้นตอนที่ 9: การเสิร์ฟอาหาร ขั้นตอนที่ 10: ความคิดเห็น </s>

# Parameters

order: prop1,prop2,prop3 Convert JSON Schema

|                                    |                                    |
|------------------------------------|------------------------------------|
| Predictions                        | Temperature                        |
| <input type="range" value="400"/>  | <input type="range" value="0.7"/>  |
| Penalize repeat sequence           | Consider N tokens for penalize     |
| <input type="range" value="1.18"/> | <input type="range" value="256"/>  |
| Penalize repetition of newlines    | Top-K sampling                     |
| <input type="checkbox"/>           | <input type="range" value="40"/>   |
| Top-P sampling                     | Min-P sampling                     |
| <input type="range" value="0.95"/> | <input type="range" value="0.05"/> |

**User:** สูตรทำส้มตำปูปลาร้า

**Llama:** นี่คือนสูตรสำหรับทำส้มตำปูปลาร้า

ส่วนผสม:

- มะละกอดิบ 1 ถ้วย
- ปูเค็ม 2 ตัว
- น้ำปลาร้า 3 ช้อนโต๊ะ
- พริกขี้หนู 5 เม็ด
- กระเทียม 2 กลีบ
- กุ้งแห้ง 1/4 ถ้วย
- ถั่วลิสงคั่ว 1/4 ถ้วย
- มะนาว 2 ลูก
- เกลือ 1 ช้อนชา
- หรือมากกว่าตามรสชาติ
- ผักชีสำหรับโรยหน้า
- กะละมังสำหรับใส่ส้มตำ

วิธีทำ:

1. เตรียมส่วนผสมทั้งหมดให้พร้อม ได้แก่ มะละกอดิบ ปูเค็ม น้ำปลาร้า พริกขี้หนู กระเทียม กุ้งแห้ง ถั่วลิสงคั่ว มะนาว เกลือ และผักชี
2. ใส่มะละกอดิบลงในกะละมัง แล้วใช้มือเปล่าบีบให้เนื้อแตกออกมาเล็กน้อย เพื่อจะได้รสชาติที่แซ่บยิ่งขึ้น
3. เติมน้ำปูเค็มประมาณ 1-2 ตัว ตามความชอบ
4. เติมน้ำปลาร้าประมาณ 3 ช้อนโต๊ะ หรือมากกว่าตามรสชาติ
5. เติมพริกขี้หนูประมาณ 5 เม็ดหรือมากกว่าตามรสชาติ
6. เติมกระเทียมสับประมาณ 2 กลีบ
7. เติมกุ้งแห้งและถั่วลิสงคั่วอย่างละ 1/4 ถ้วย
8. บีบน้ำมะนาวสด 2 ลูกลงไปลงในกะละมัง
9. เติมเกลือเล็กน้อยเพื่อดึงรสชาติให้กลมกล่อมยิ่งขึ้น
10. ใช้มือคลุกเคล้าส่วนผสมทั้งหมดเข้าด้วยกันจนเข้ากันดี
11. ชิมและปรับรสตามที่ชอบ
12. ตักส้มตำปูปลาร้าใส่จานและ

**User:** และอะไรต่อ

**Llama:** โรยหน้าด้วยผักชีสด

# RAG: Retrieval Augmented Generation

