Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

CIS 6930 Special Topics in Large Language Models

Retrieval Augmented Language Model

Outline

- Part 1: Introduction to Retrieval Augmentation
- Part 2: Retrieval Augmentation Architectures (main body)
- Part 3: Other Interesting Questions of Retrieval Augmentation
- Part 4: Future (more open questions)

Introduction to Retrieval Augmentation

Traditional Language Model

S = Where are we going

$\underbrace{\text{Where are we}}_{\text{Previous words (context)}}$

Previous words (context)

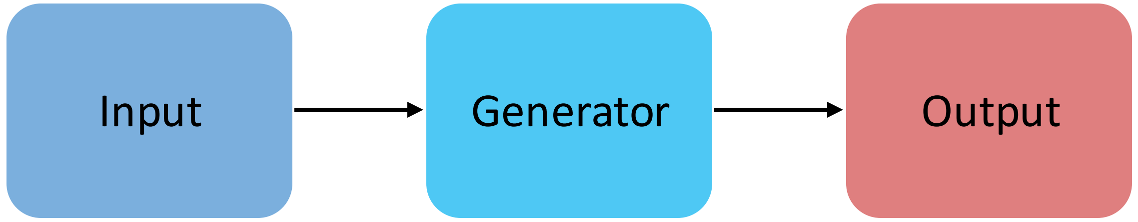
Word being predicted

P(S) = P(Where) * P(are | Where) * P(we | Where are) * P(going | Where are we)

Training objective: maximize the joint probability of the observed text

Next Word Prediction

Plain vanilla sequence-to-token



Problems:

- Lack user interface

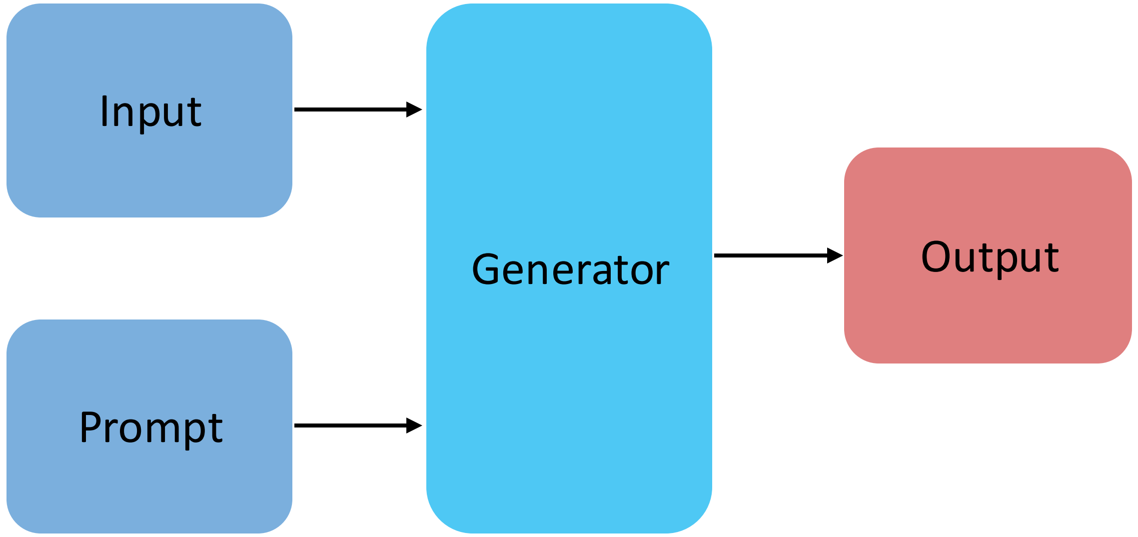
Solutions:

- Prompt your model
- Instruction tune your model to follow prompts
- Align your model with human preferences

Prompt-Tuned Language Model

Problems:

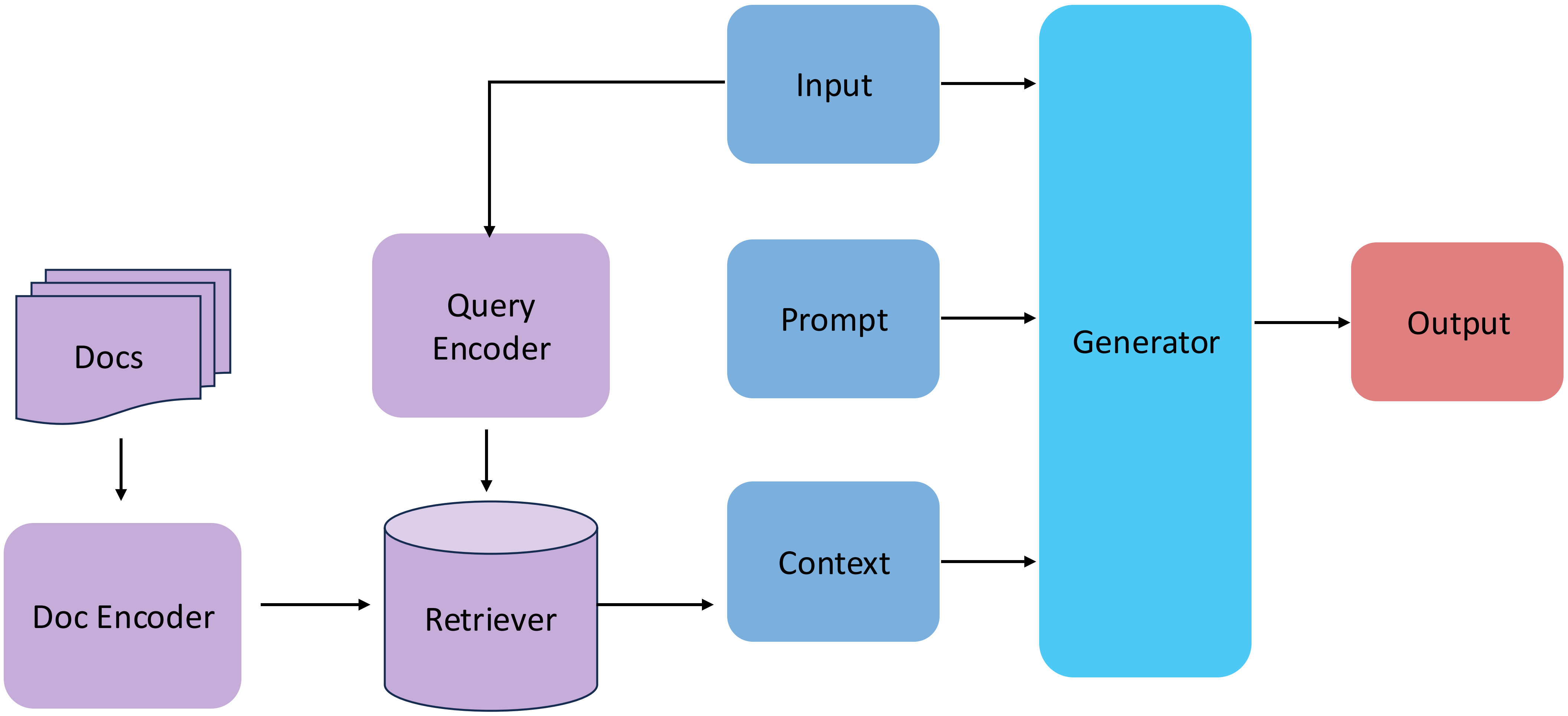
- Hallucination
- Attribution
- Staleness
- Revisions
- Customization



Solutions:

- Contextualized to external memory (RAG)

Contextualization



Two Paradigms

- Closed book (knowledge in parameters) vs Open book (external source)

- Parametric vs Non-parametric / Semi-parametric

Why does RAG solve the issues?

- Choosing contextualized documents allows **customization**, which means you can **revise** knowledge and don't suffer from **staleness**

- Grounding means you have less **hallucinations**, and you can do citations and **attribution** by pointing back to the source

Retrieval Augmentation Architectures

Designs of Retrieval Augmentation Architectures

- What happens during training?

Update the generator (LM)? Update the query encoder? Update the document encoder? Update all? Pretrain from scratch or not?

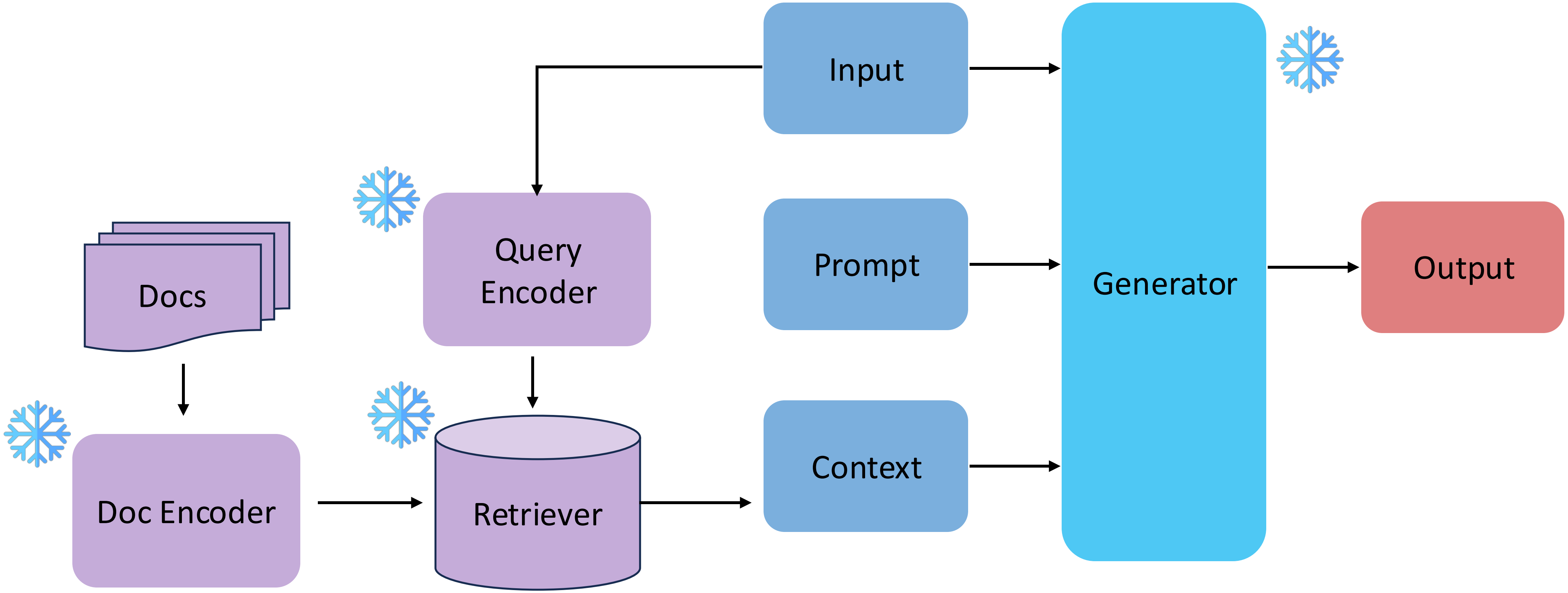
- What happens during inference?

Different or same retrieved documents? etc.

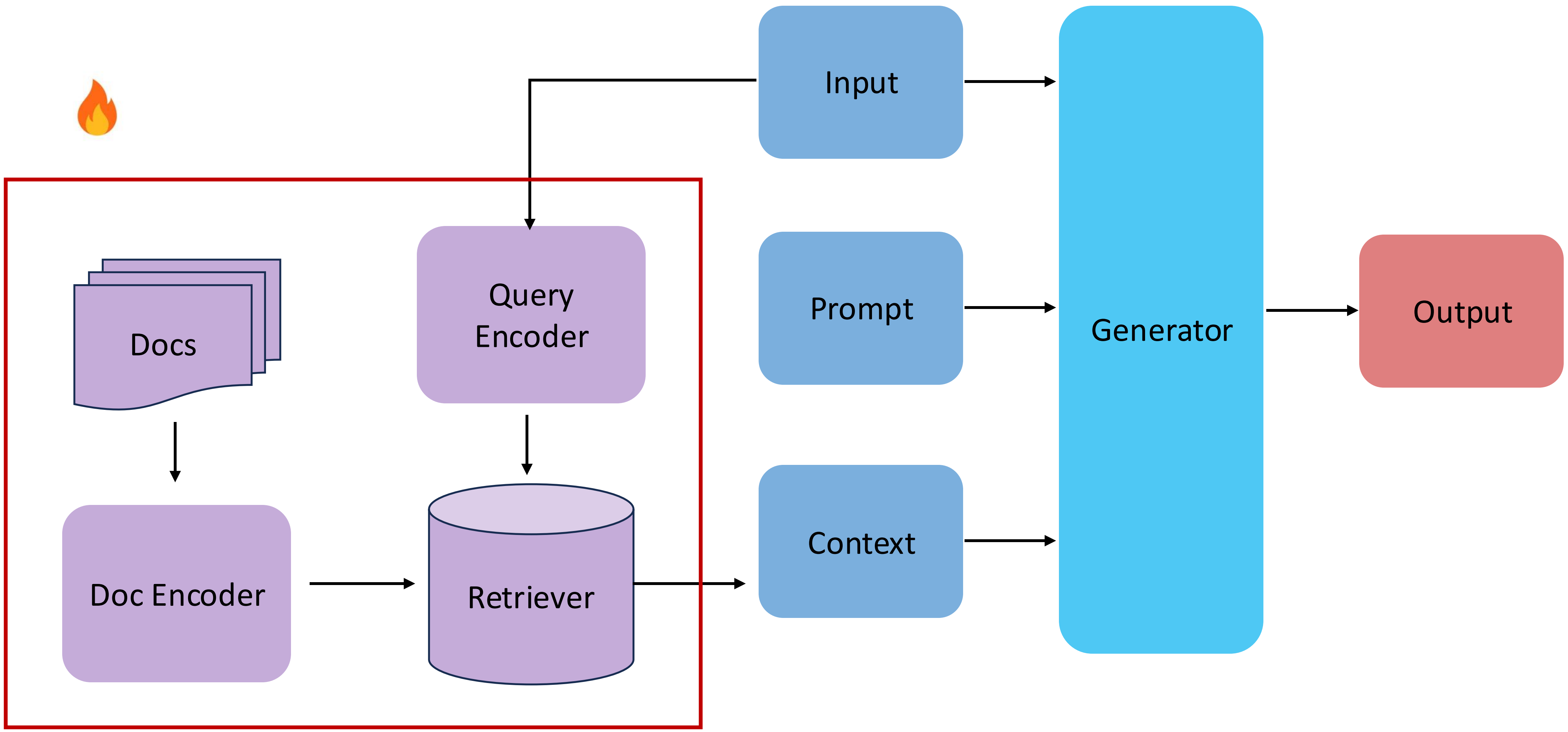
I. Frozen RAG

- No training, in-context learning only
- Everything frozen

LM prompts is hand-tuned to maximize in-context learning performance



II. Contextualization via Retrieval



Sparse Retrieval

- TF-IDF and BM25 (variant of TF-IDF) (Robertson, Sparck-Jones et al)

$$\mathrm{TFIDF}(Q,d) = \Sigma_{t \in Q} \frac{\mathrm{tf}_{t,d}}{l_d} \cdot \mathrm{idf}_t.$$

$$\Sigma_{t \in Q} \frac{\mathrm{tf}_{t,d} \cdot (k+1)}{\mathrm{tf}_{t,d} + k \cdot \left(1 - b + b \cdot \frac{l_d}{\mathrm{mean}(l_d)}\right)} \cdot \ln\left(1 + \frac{N - \mathrm{df}_t + 0.5}{\mathrm{df}_t + 0.5}\right).$$

- Used in DrQA (Chen et al., 2017)

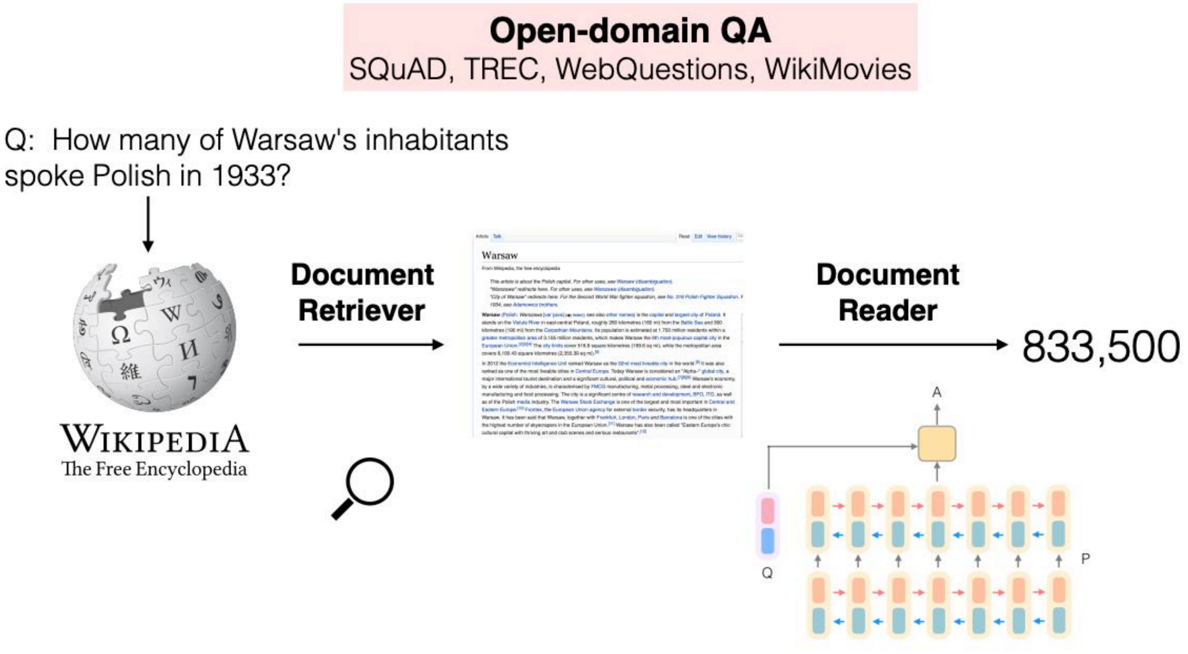
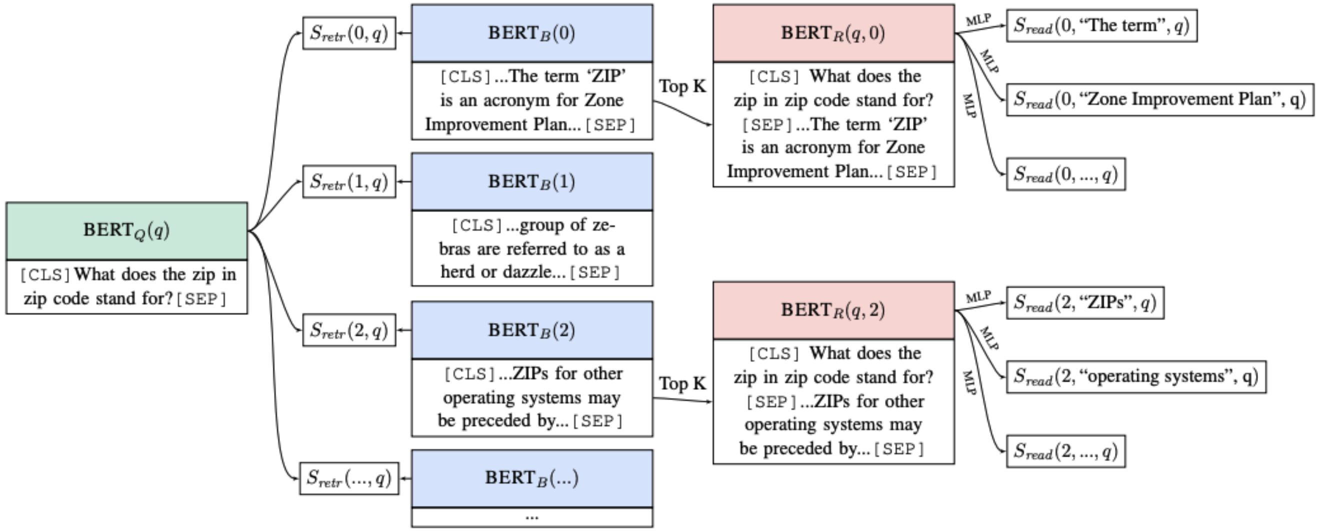
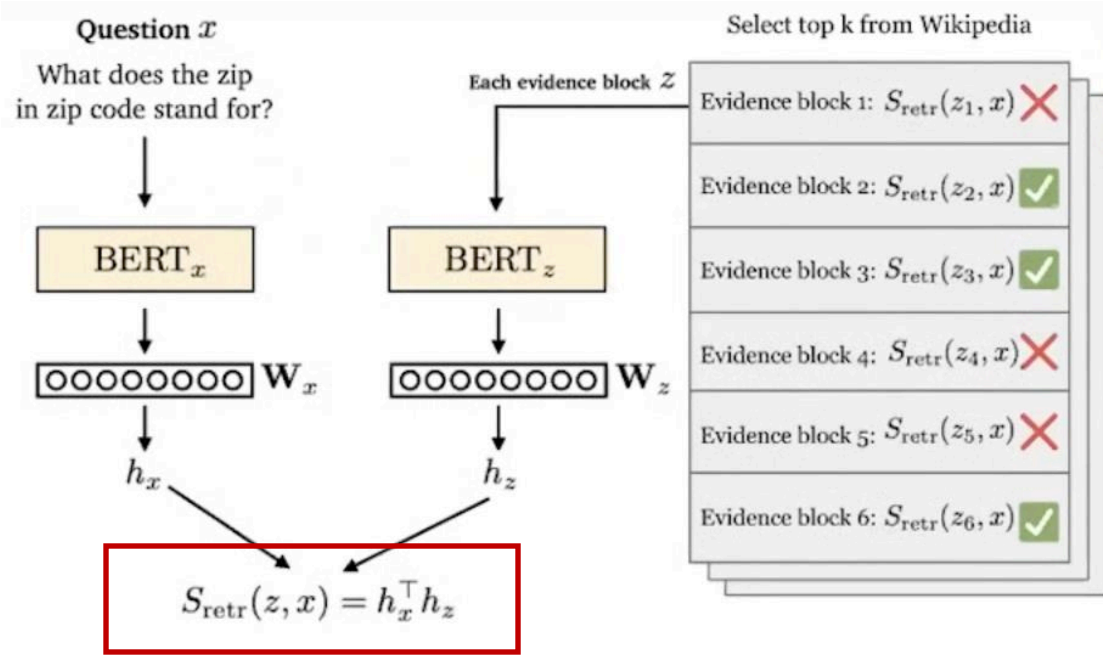


Figure 1: An overview of our question answering system DrQA.

Dense Retrieval

- OrQA (Lee et al., 2019)
- Dense Passage Retriever (Karpukhin, Oguz et al., 2020)

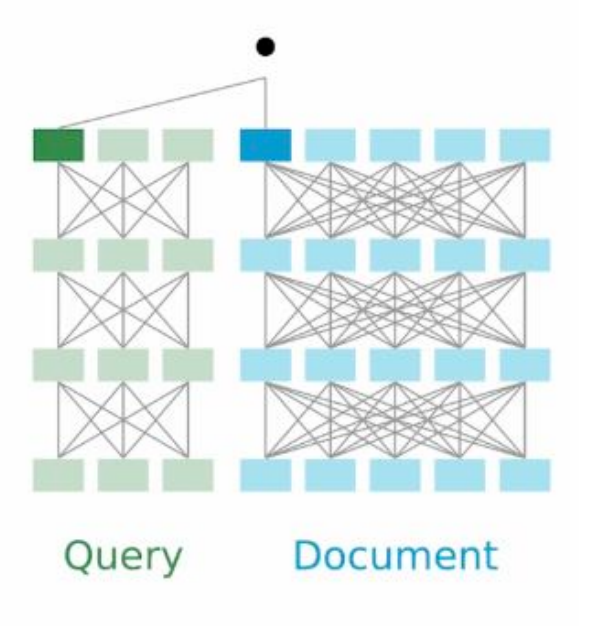




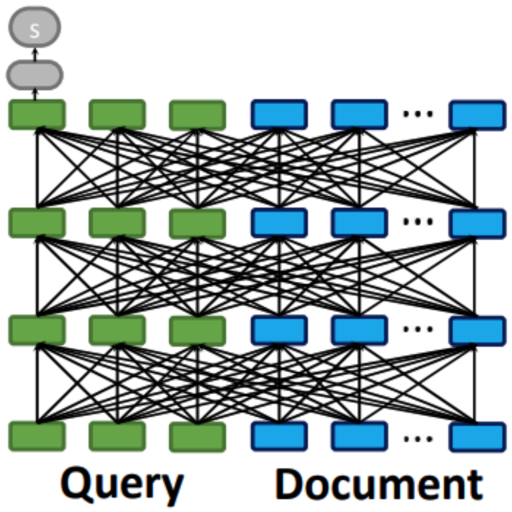
Dot Product –
semantic similarity

Dense Retrieval beyond Dot Product

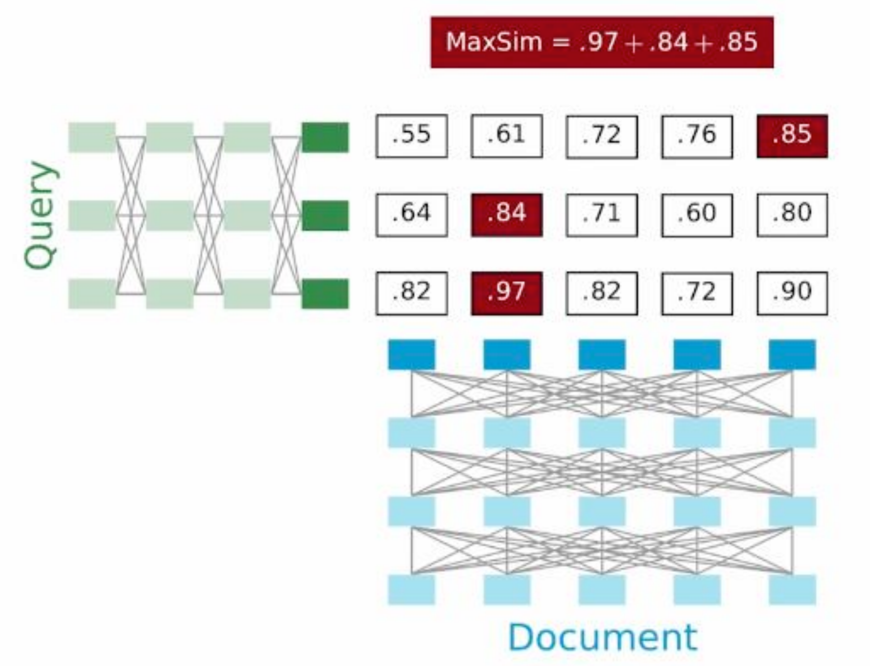
- ColBERT (Khattab et al., 2020) – how do query and documents interact



Separate encoders
NN to learn similarity



Cross-encoders
all-to-all interactions



ColBERT (late interaction)

Sparse Retrieval vs Dense Retrieval

- Sparse Representation (lexical)

Corresponding actual specific words

Easier to interpret how documents are ranked by a given query

Low cost for building new sparse search engine infrastructure

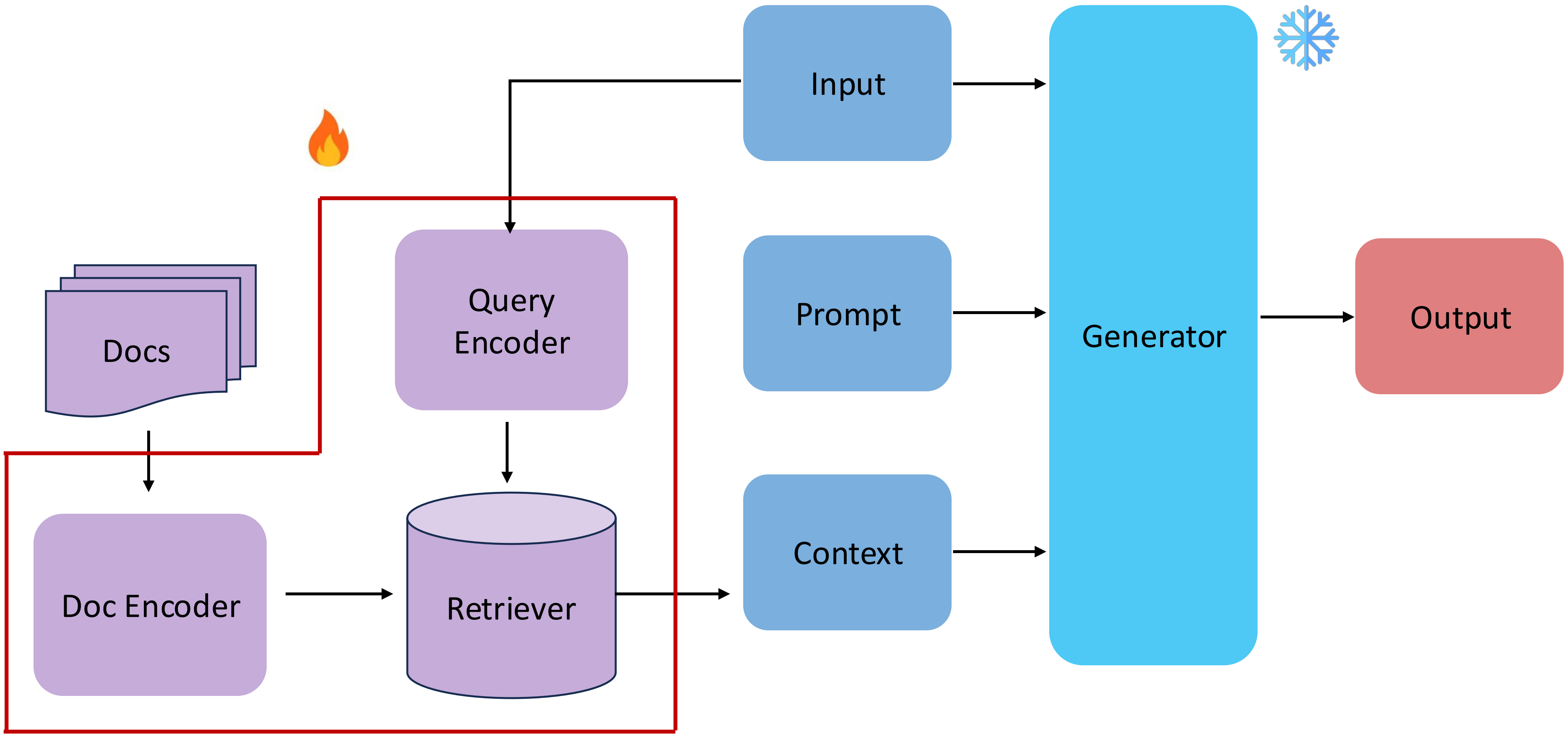
- Dense Representation (semantic)

Contextualized representation

Empirically better performances

Easier to scale for efficient implementation

III. Contextualizing the Retriever for the Generator



Generator as a frozen black-box LM

- RePlug (Shi et al., 2023) – inference

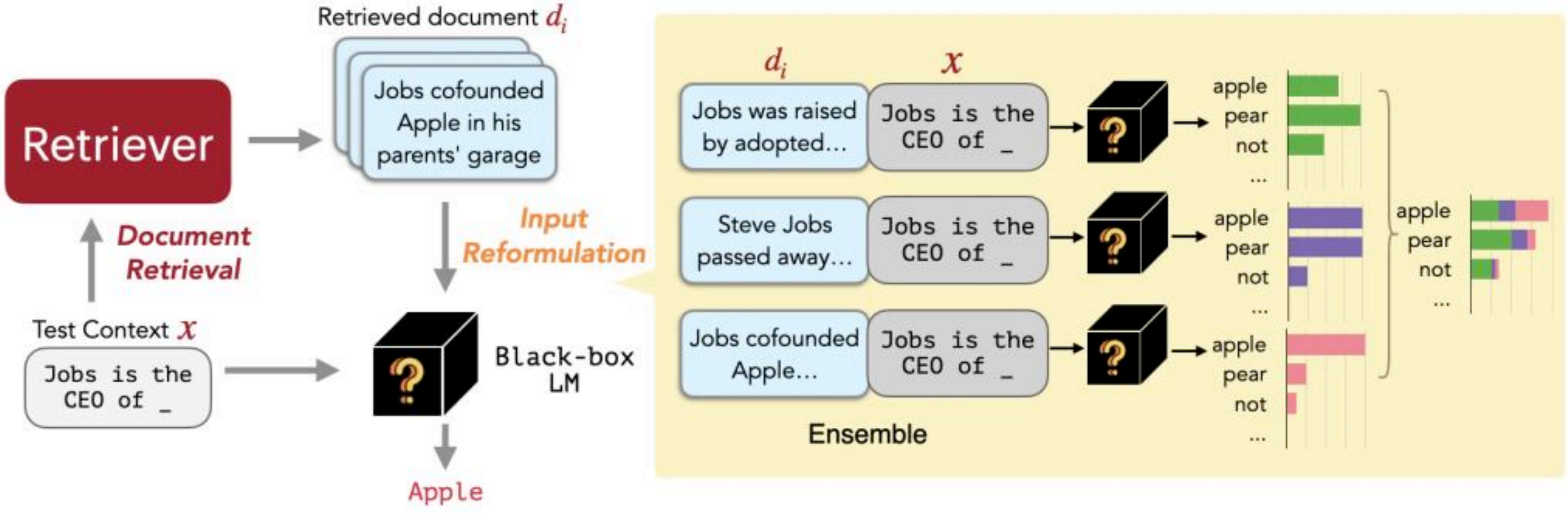


Figure 2: **REPLUG at inference** (§3). Given an input context, REPLUG first retrieves a small set of relevant documents from an external corpus using a retriever (§3.1 *Document Retrieval*). Then it prepends each document separately to the input context and ensembles output probabilities from different passes (§3.2 *Input Reformulation*).

Generator as a frozen black-box LM

- RePlug (Shi et al., 2023) – training

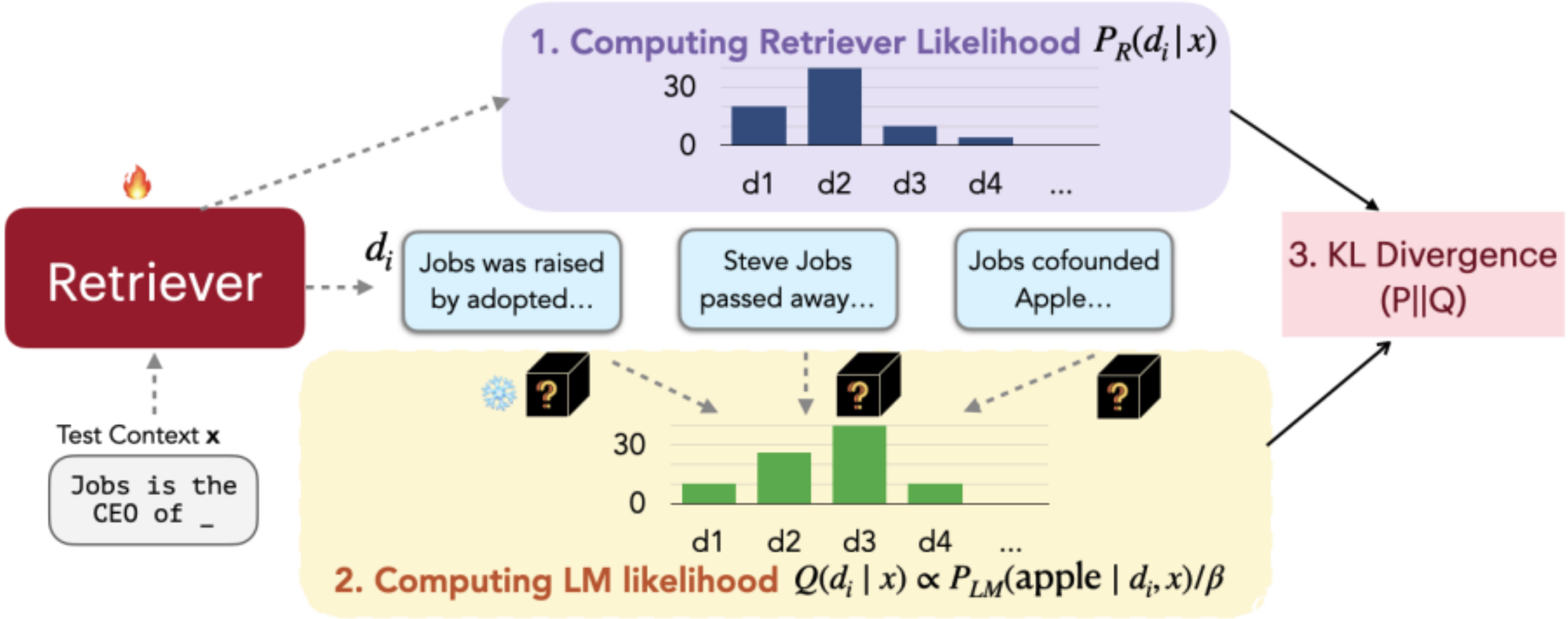
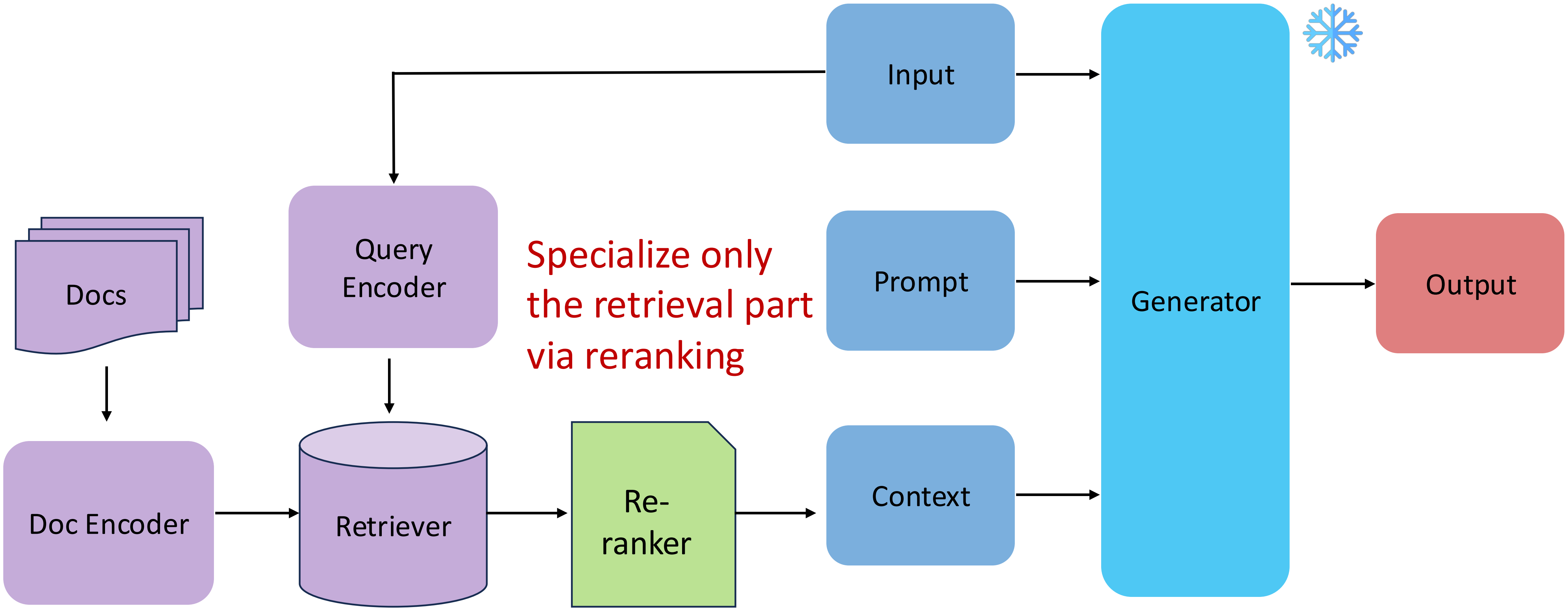
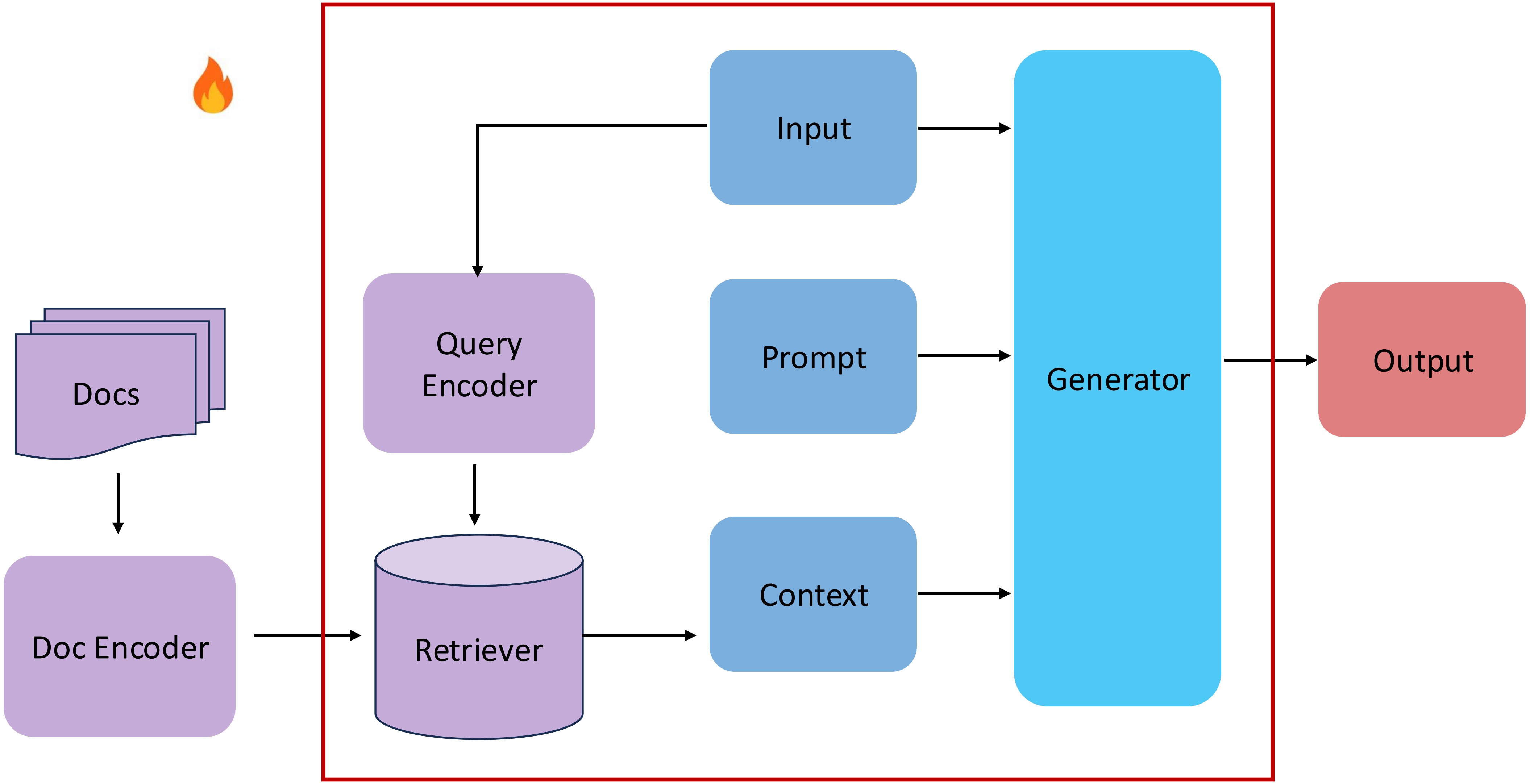


Figure 3: **REPLUG LSR training process (§4).** The retriever is trained using the output of a frozen language model as supervision signals.

IV. Contextualization via retrieve-rerank



V. Contextualization of both



Fine-tune both generator and retriever

- RAG (Lewis et al., 2020)

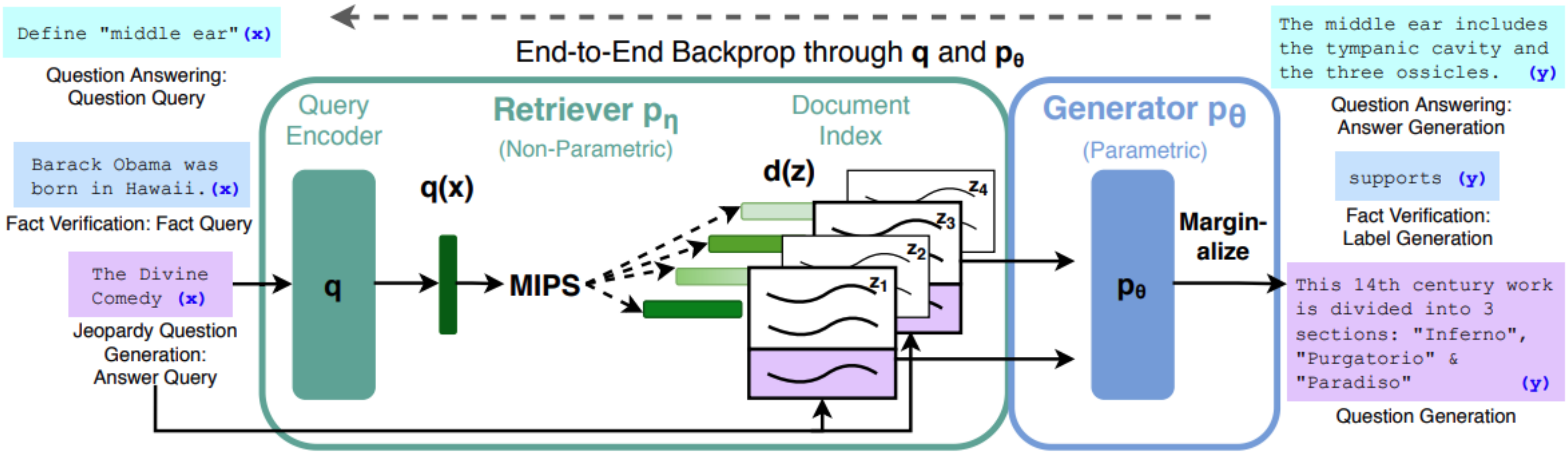
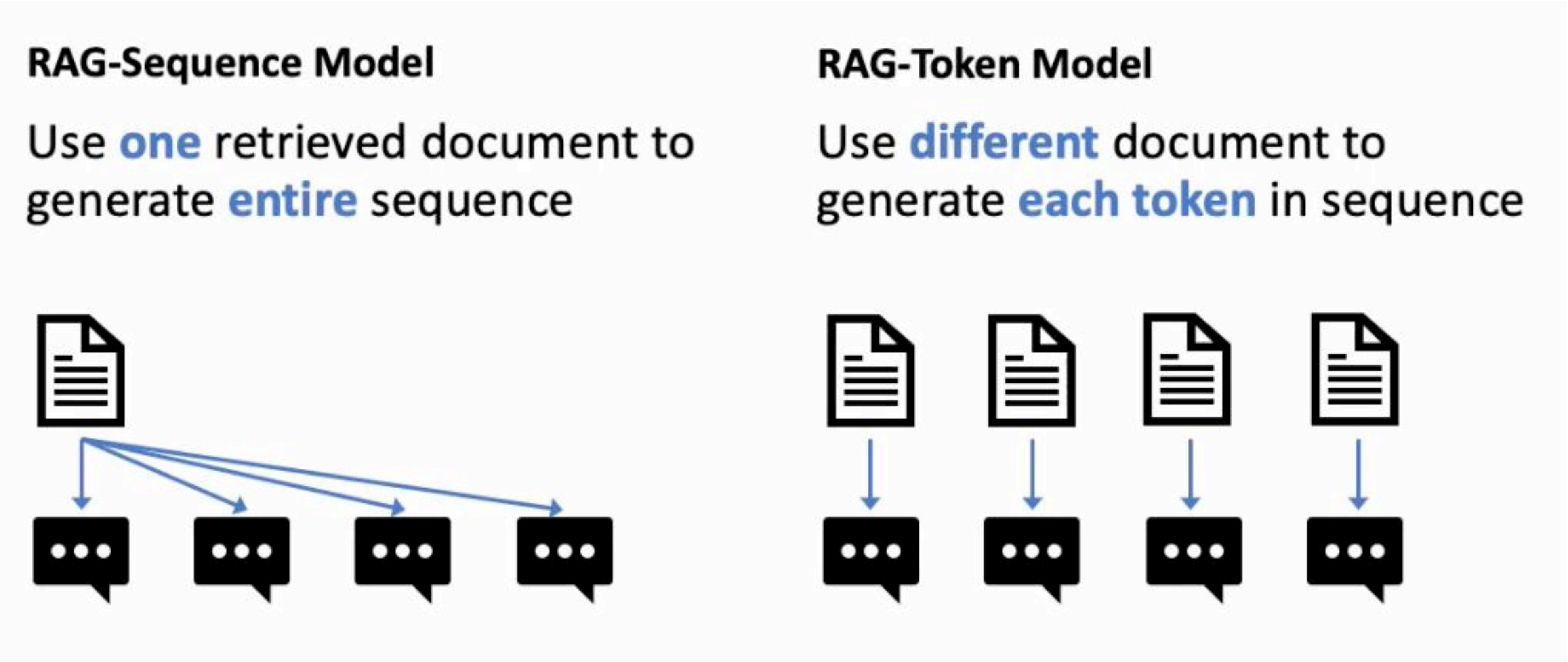


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Two Types of RAG

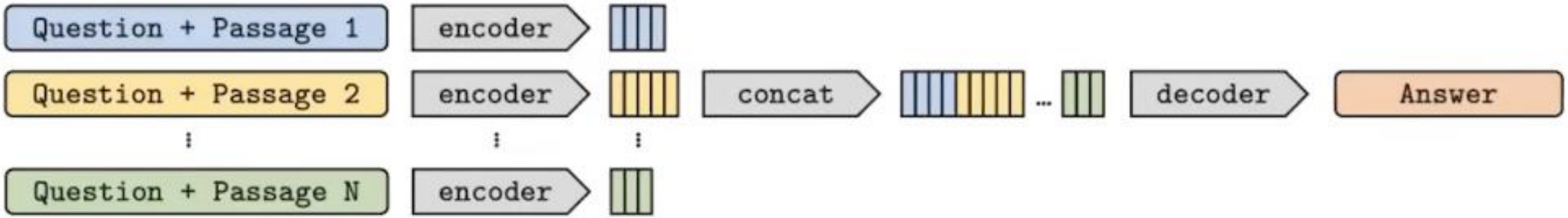
- RAG (Lewis et al., 2020)



Fusion in the Decoder – increase k

- FiD (Izacard & Grave 2020)

Address the limitation of small k in RAG – fusion in the decoder directly



Generator with kNN-based retriever

- kNN-LM (Khandelwal et al., 2019)

late interpolation of parametric LM and non-parametric kNN retriever

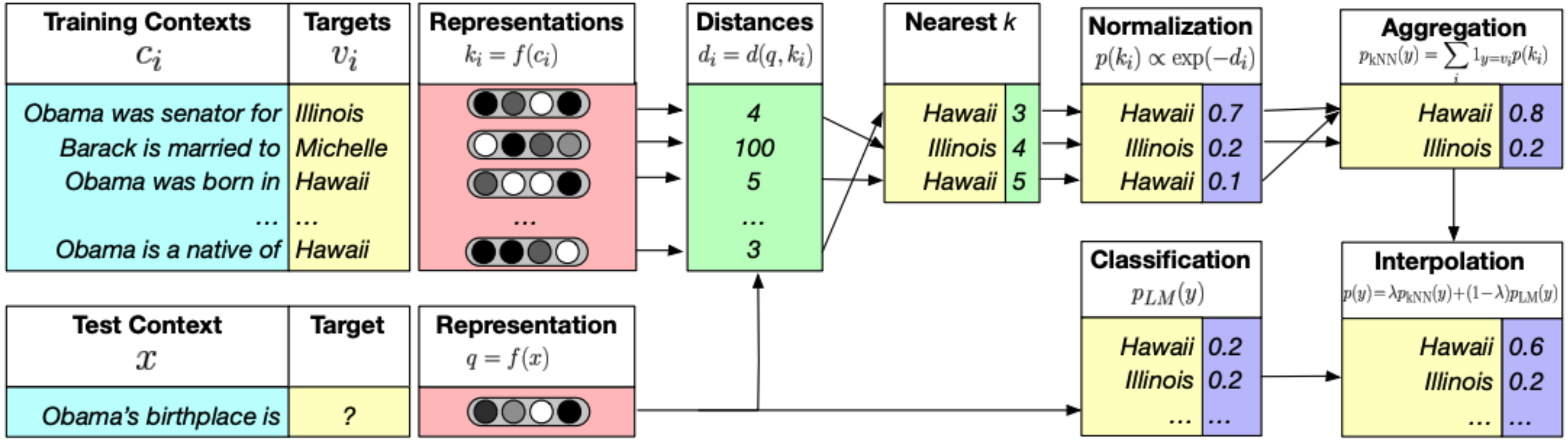
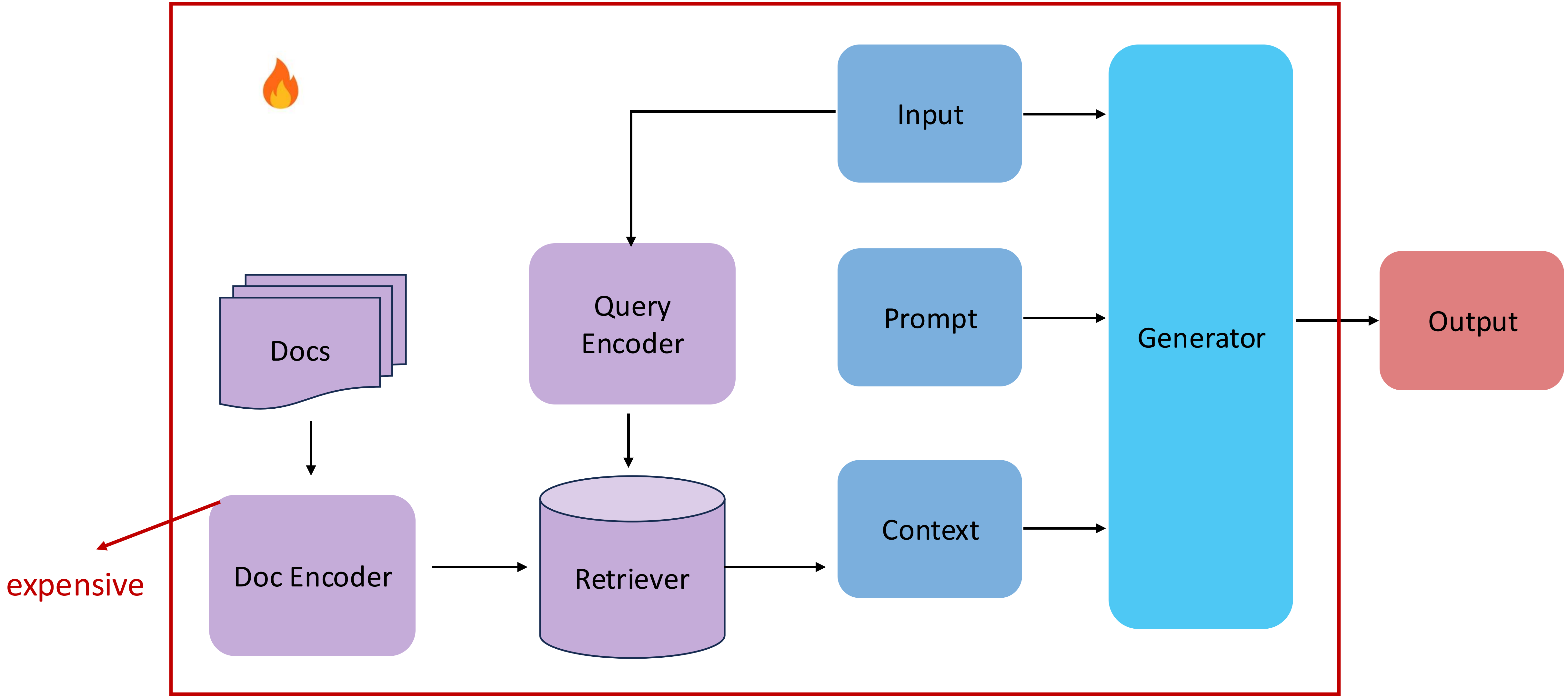


Figure 1: An illustration of kNN-LM. A datastore is constructed with an entry for each training set token, and an encoding of its leftward context. For inference, a test context is encoded, and the k most similar training contexts are retrieved from the datastore, along with the corresponding targets. A distribution over targets is computed based on the distance of the corresponding context from the test context. This distribution is then interpolated with the original model's output distribution.

VI. Contextualization all the way

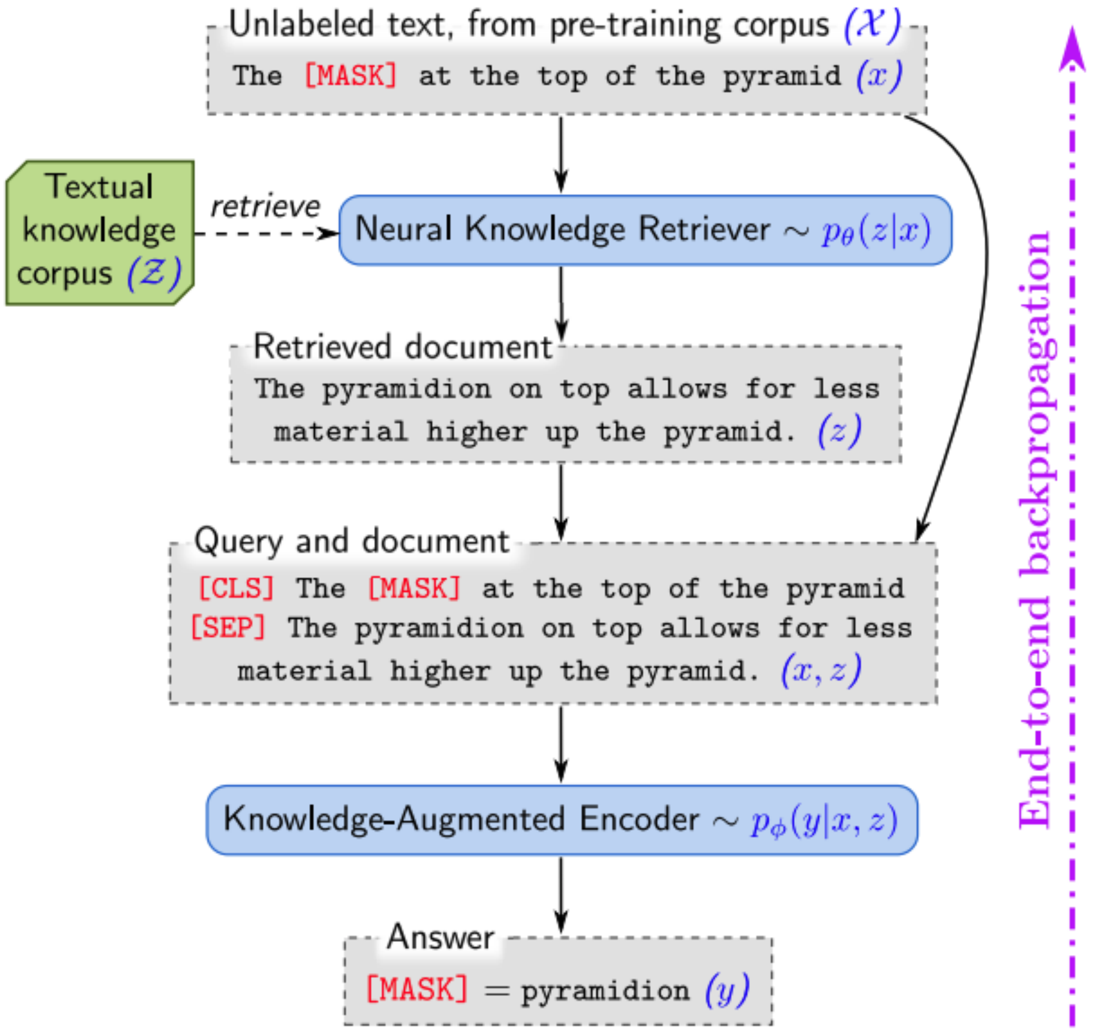


Backpropagate all the way

- REALM (Guu et al., 2020)

A classical work of non-frozen retrieval augmented LMs

Signal from language modeling objective backpropagates all the way through the retriever, query encoder, and document encoder



Other Interesting Questions of Retrieval Augment

Other Interesting Questions

- When to retrieve?
- Legal risk of training or retrieval data source?
- Does the order of retrieved documents matter?
- Extension of retrieval augmentation?
- Combine with instruction tuning?
- Multimodal RAG?

When to Retrieve

- FLARE (Jiang, Xu, Gao, Sun et al., 2023)

LM will decide when to retrieve and when not

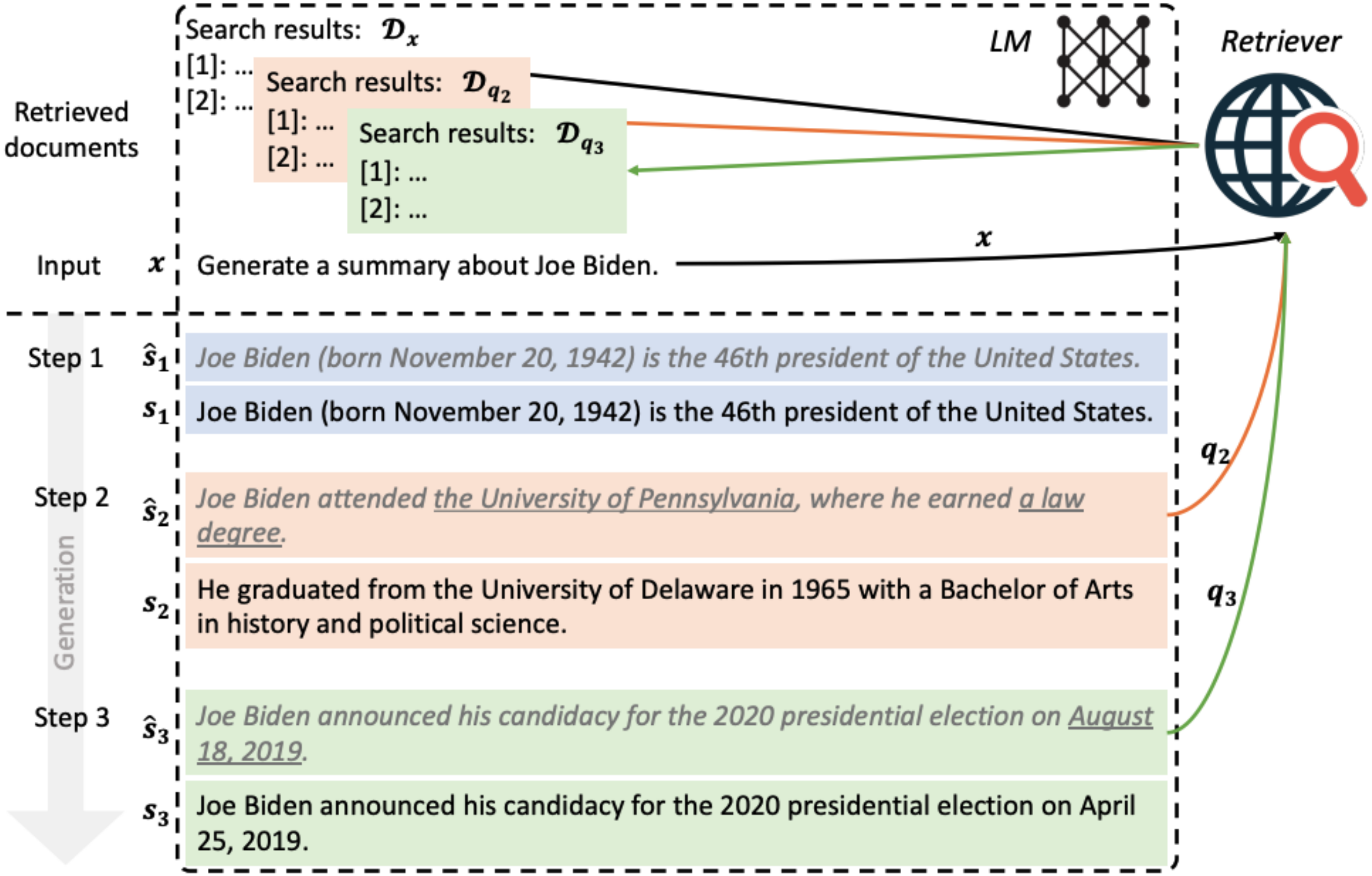
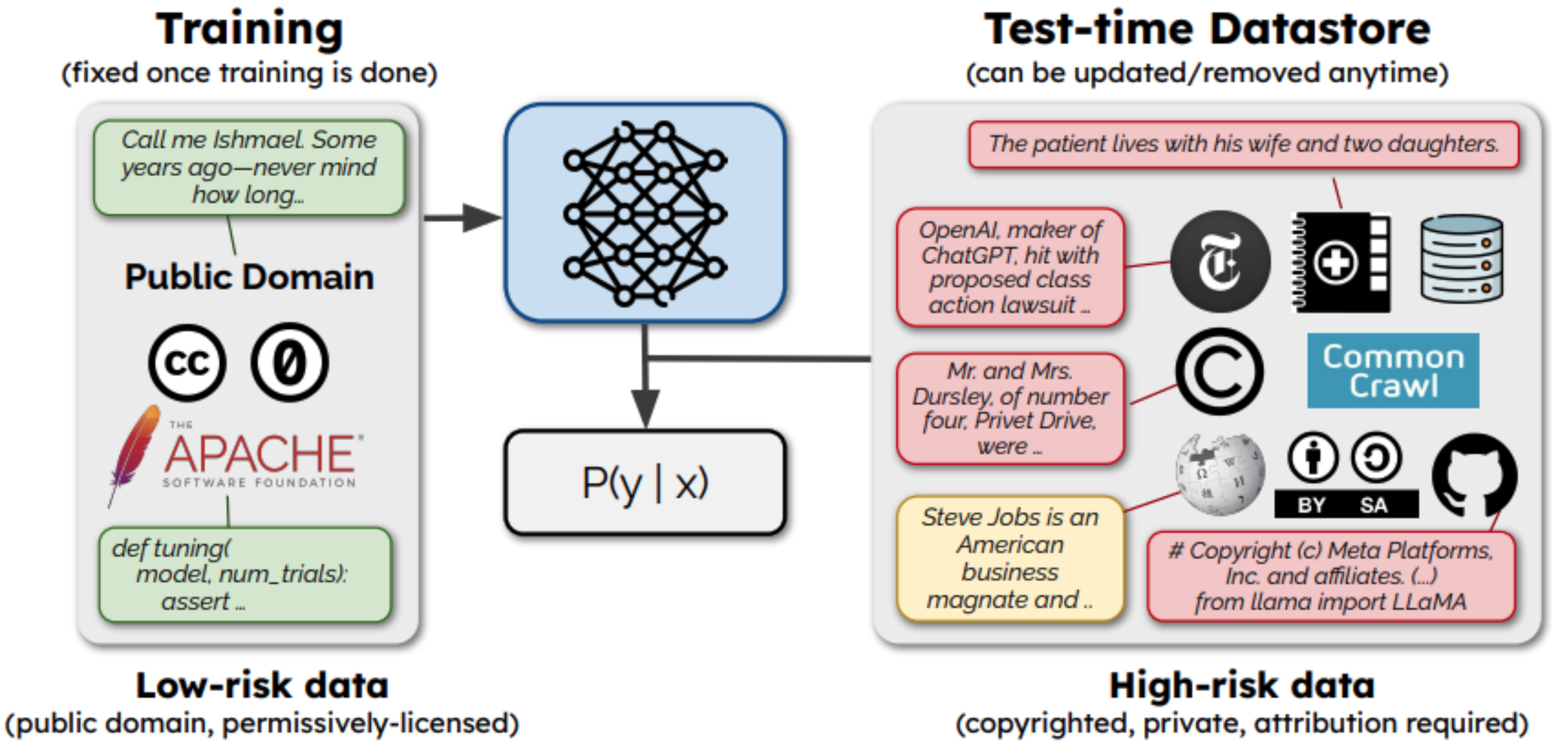


Figure 1: An illustration of forward-looking active retrieval augmented generation (FLARE). Starting with the user input x and initial retrieval results $\mathcal{D}_x$, FLARE iteratively generates a temporary next sentence (shown in *gray italic*) and check whether it contains low-probability tokens (indicated with underline). If so (step 2 and 3), the system retrieves relevant documents and regenerates the sentence.

Isolating legal risk with retrieval

- SILO (Min, Gururangan et al., 2023)

Parametric LM under training



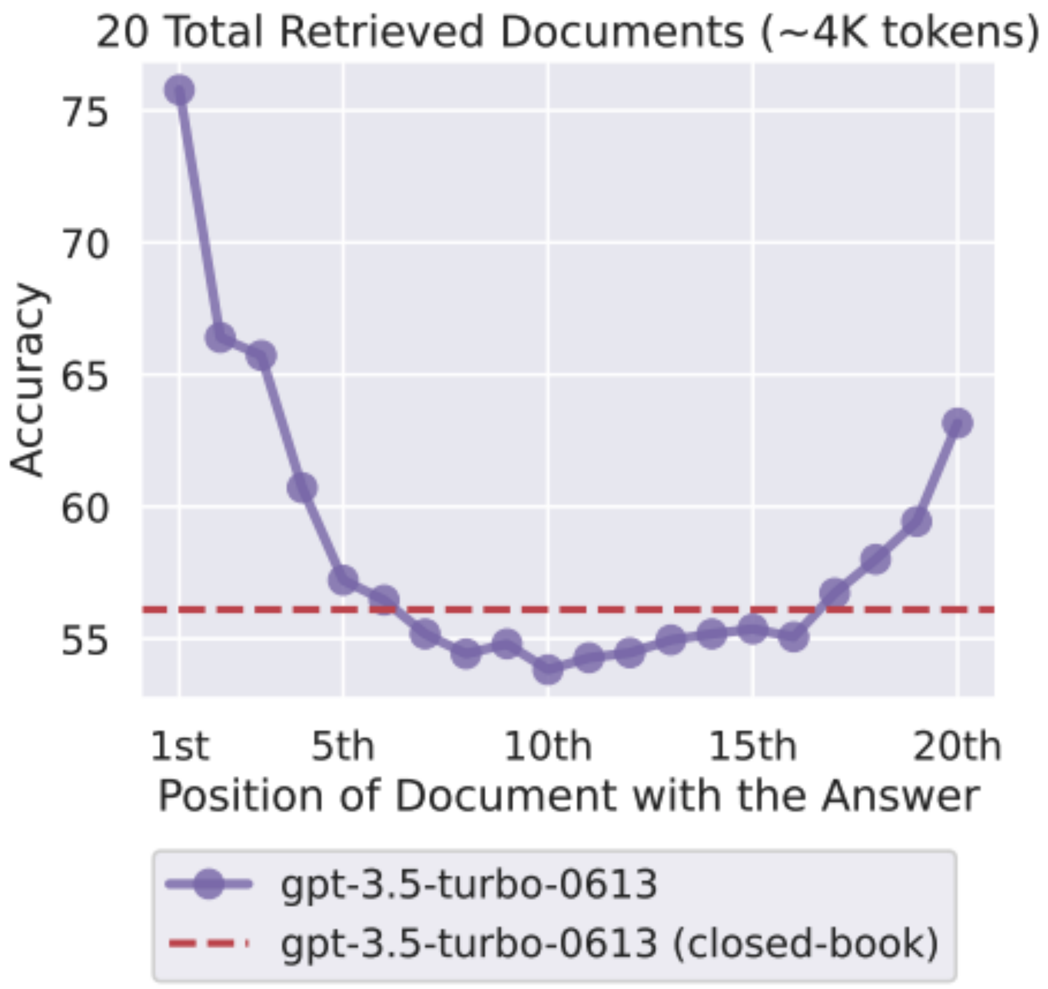
Non-parametric data store for testing-time retrieval

The order of retrieved documents

- Liu et al., 2023

lost in the middle when LM use long contexts

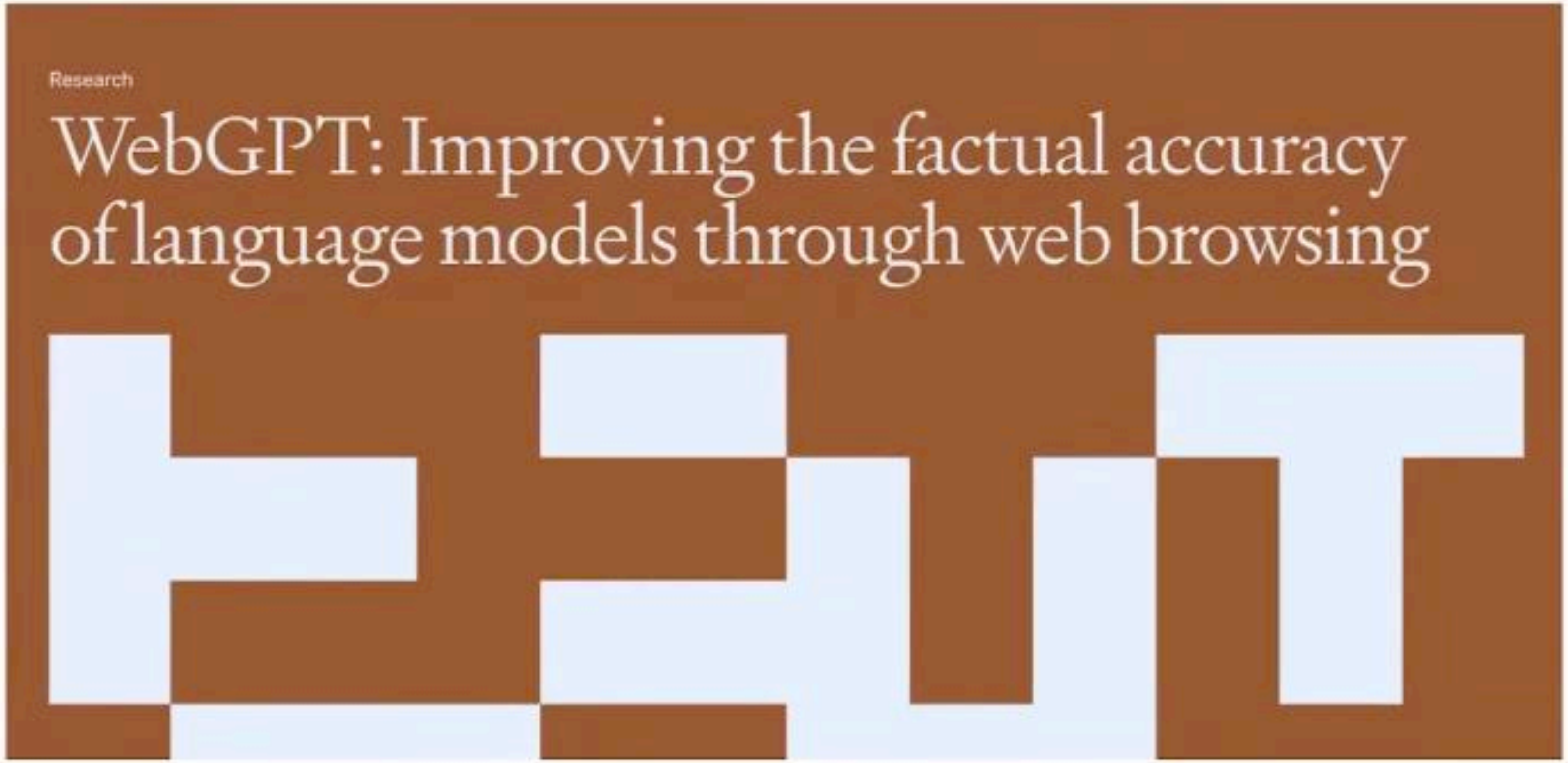
LM attend more to beginning and latter tokens, but less to the middle



Extension of Retrieval Augmentation

- WebGPT (Nakano et al., 2021)

The retrieved documents can be replaced by anything



Extension of Retrieval Augmentation

- Toolformer (Shick et al., 2021)

Can be generalized to all kinds of tools

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick Jane Dwivedi-Yu Roberto Dessì[†] Roberta Raileanu
Maria Lomeli Luke Zettlemoyer Nicola Cancedda Thomas Scialom
Meta AI Research [†]Universitat Pompeu Fabra

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Figure 1: Exemplary predictions of Toolformer. The model autonomously decides to call different APIs (from top to bottom: a question answering system, a calculator, a machine translation system, and a Wikipedia search engine) to obtain information that is useful for completing a piece of text.

Combined with Instruction Tuning

- InstructRetro (Wang et al., 2023)
- RA-DIT (Lin, Chen et al, 2023)

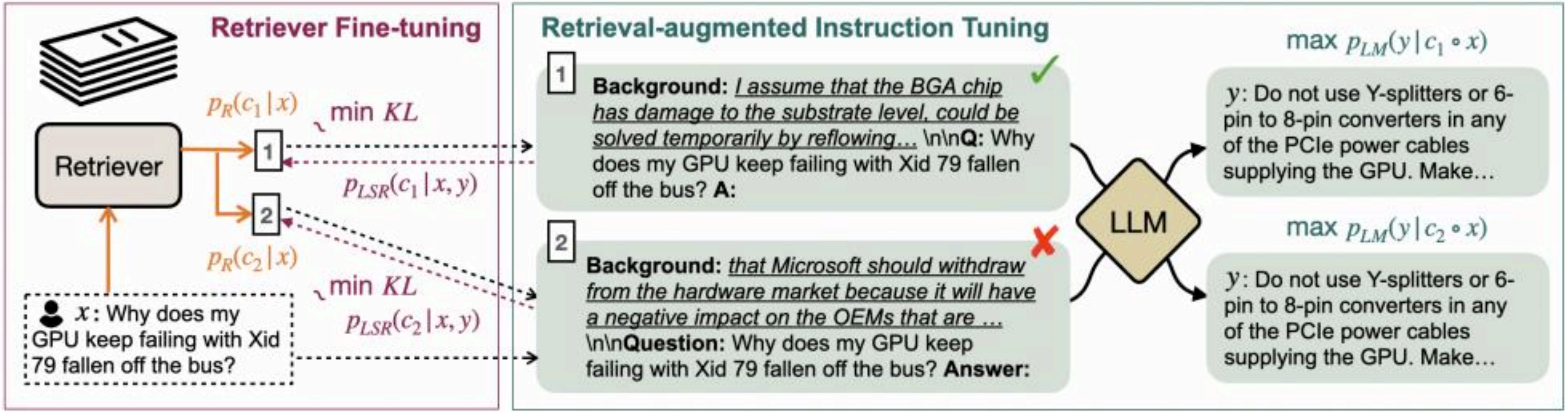
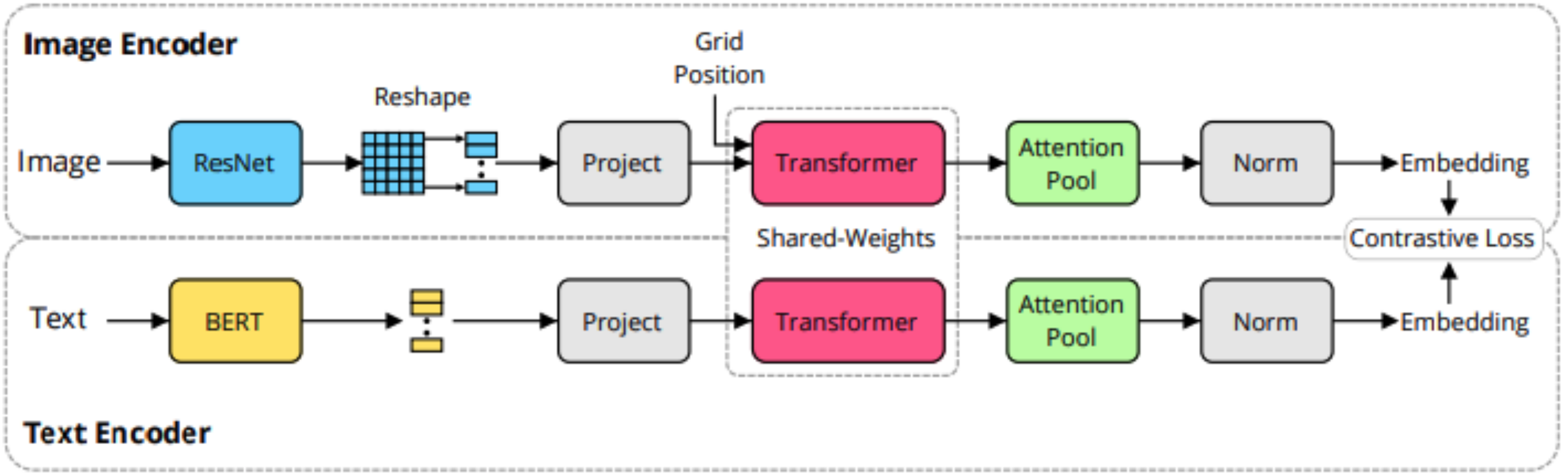
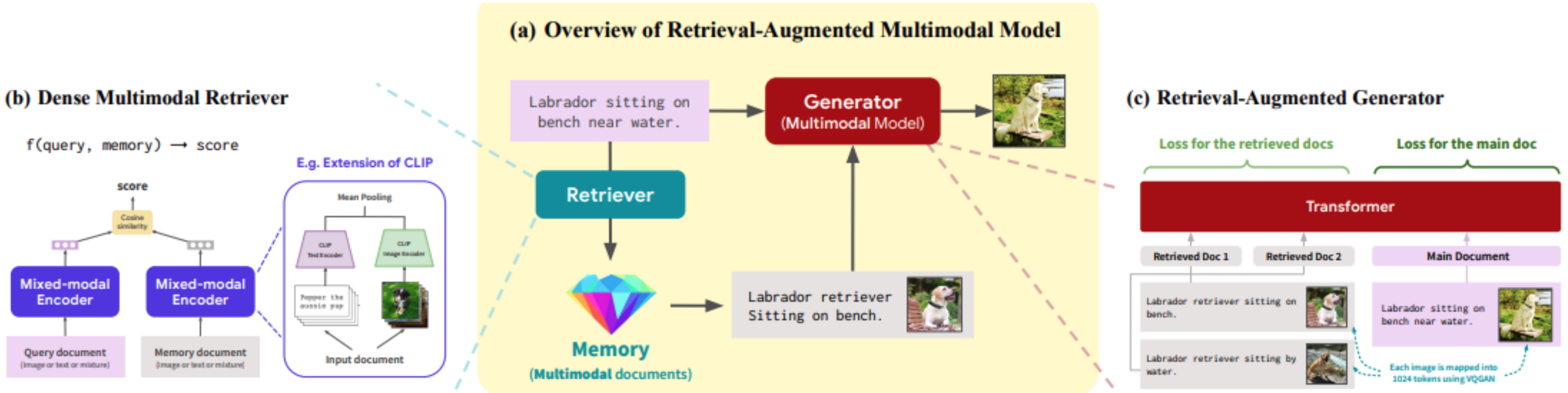


Figure 1: The RA-DIT approach separately fine-tunes the LLM and the retriever. For a given example, the LM-ft component updates the LLM to maximize the likelihood of the correct answer given the retrieval-augmented instructions (§2.3); the R-ft component updates the retriever to minimize the KL-Divergence between the retriever score distribution and the LLM preference (§2.4)

Multimodal RAG

- Gur et al., 2021
- Yasunaga et al, 2023





Future – more open questions

More open questions

- Joint from-scratch pretraining is still underexplored
- What do scaling laws look like? (Scale LM in terms of params or tokens, Scale the retriever in terms of params or chucks, Scale the index size during inference)
- Can we fully decouple memorization from generalization, decouple knowledge from generation?
- Are there smart ways to create synthetic data for RAG?
- How do we properly evaluate RAG system?
- Zero-shot domain generalization

Thank you!

UF | Herbert Wertheim College of Engineering
UNIVERSITY *of* FLORIDA