PROVENCE

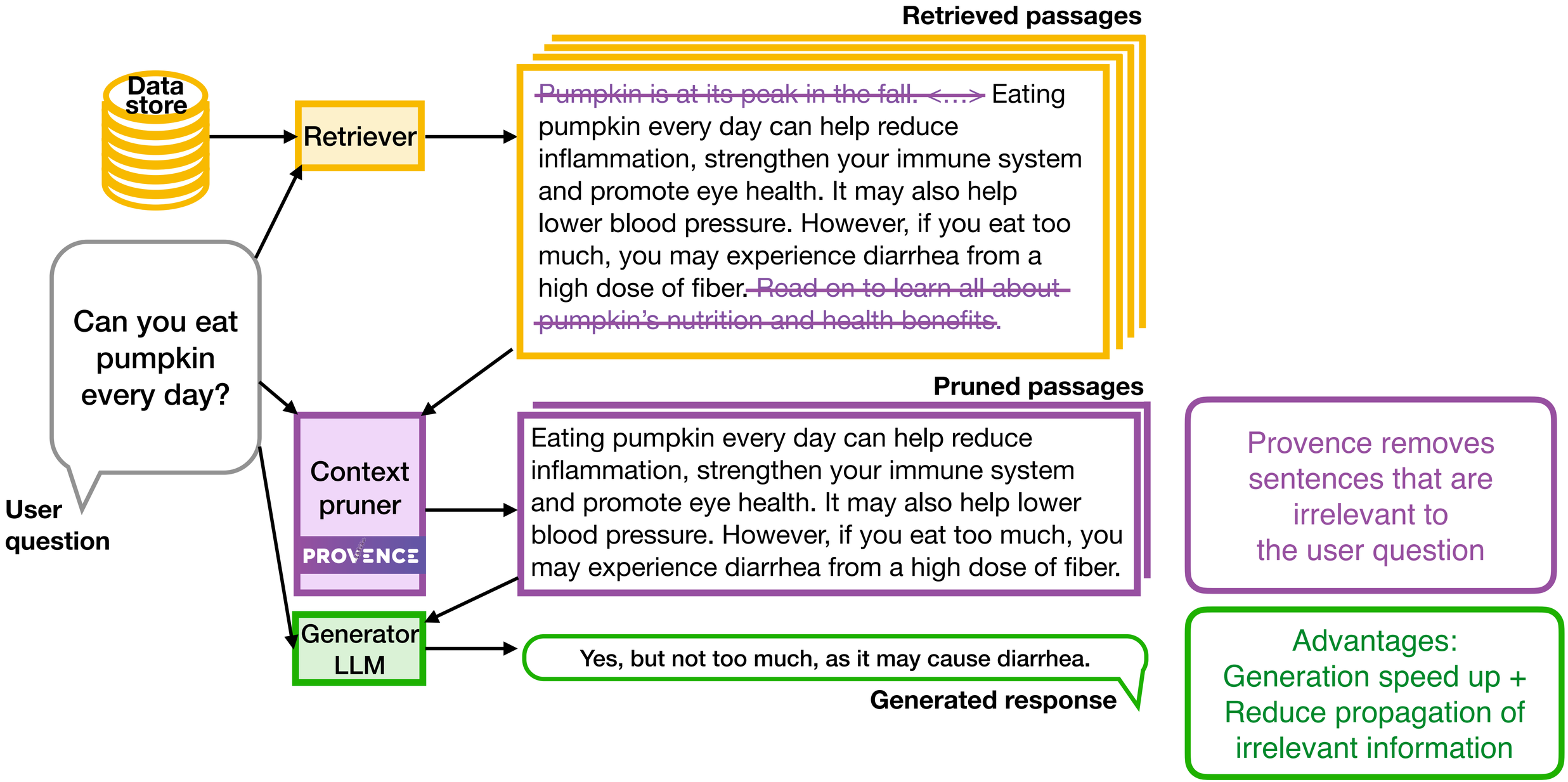
Provence: efficient and robust context pruning for retrieval-augmented generation

Nadezhda Chirkova, Thibault Formal,
Vassilina Nikoulina, Stéphane Clinchant

ICLR'25

NAVER LABS
Europe

PROVENCE Pruning and Reranking Of retrieVEd relevaNt ContExts



Existing context pruning approaches

adaptability ***efficiency*** ***robustness***

Approach	Query-dep.	Granularity	Output	Type	Base arch.	Multi-domain testing	Model release
Selective Context	No	token-level	% of tokens	extr.	Llama-7B / GPT2	Yes	Yes
LLMLingua	No	token-level	% of tokens	extr.	Alpaca-7B / GPT2	Yes	Yes
LongLLMLingua	Yes	token-level	% of tokens	extr.	Llama-2-7B-chat	Yes	Yes
LLMLingua2	No	token-level	% of tokens	extr.	RoBERTa / mBERT	Yes	Yes
RECOMP extr.	Yes	sent.-level	k sentences	extr.	BERT	No	Yes
RECOMP abstr.	Yes	sent.-level	$\geqslant 0$ sentences	abstr.	T5-L	No	Yes
FilCo	Yes	sent.-level	1 sentence	abstr.	T5-XL / Llama-2-7B	No	No
COMPACT	Yes	sent.-level	$\geqslant 0$ sentences	abstr.	Mistral-7B	No	Yes
`Provence` (ours)	Yes	sent.-level	$\geqslant 0$ sentences	extr.	DeBERTa	Yes	Yes

Violet: practical solution
Orange: less-practical solution

In Provence, we aim to train an ***adaptable***, ***efficient*** and ***robust*** context pruner ready to be used **out-of-the-box** for any question answering domain and any LLM

Provence training

Step 1: retrieve passages relevant to the train questions

Step 2: generate synthetic labels using a strong LLM

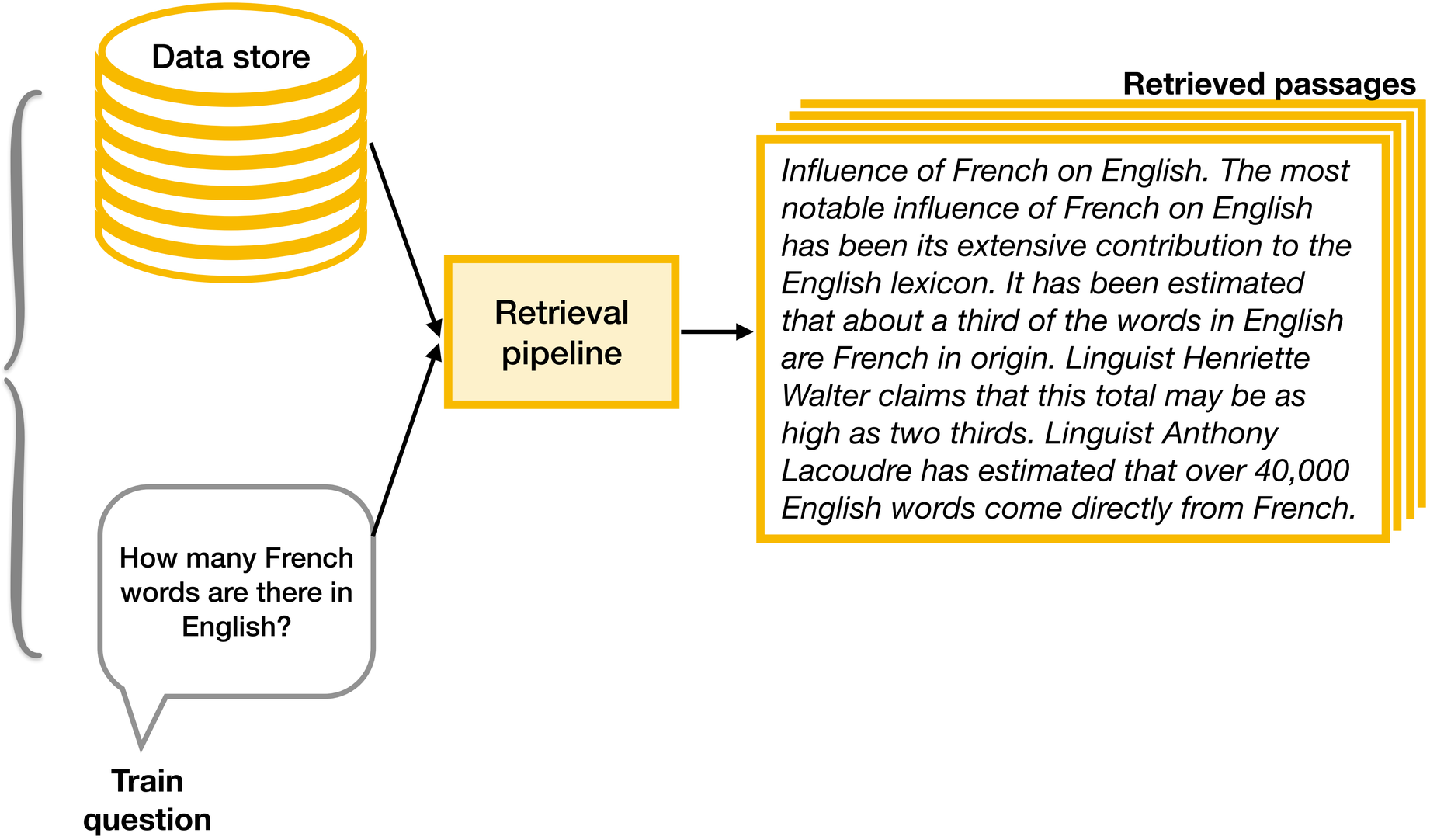
Step 3: train a compact context compressor using the synthetic labels

Step 1: retrieval for the train set

Train set

MS MARCO

- A dataset commonly used to train retrievers
- Diverse domain data (crawled from the web)



Step 2: synthetic labels generation

Retrieved passage

Influence of French on English. The most notable influence of French on English has been its extensive contribution to the English lexicon. It has been estimated that about a third of the words in English are French in origin. Linguist Henriette Walter claims that this total may be as high as two thirds. Linguist Anthony Lacoudre has estimated that over 40,000 English words come directly from French.

LLama3-8b

Data labeler LLM

Labelled passage

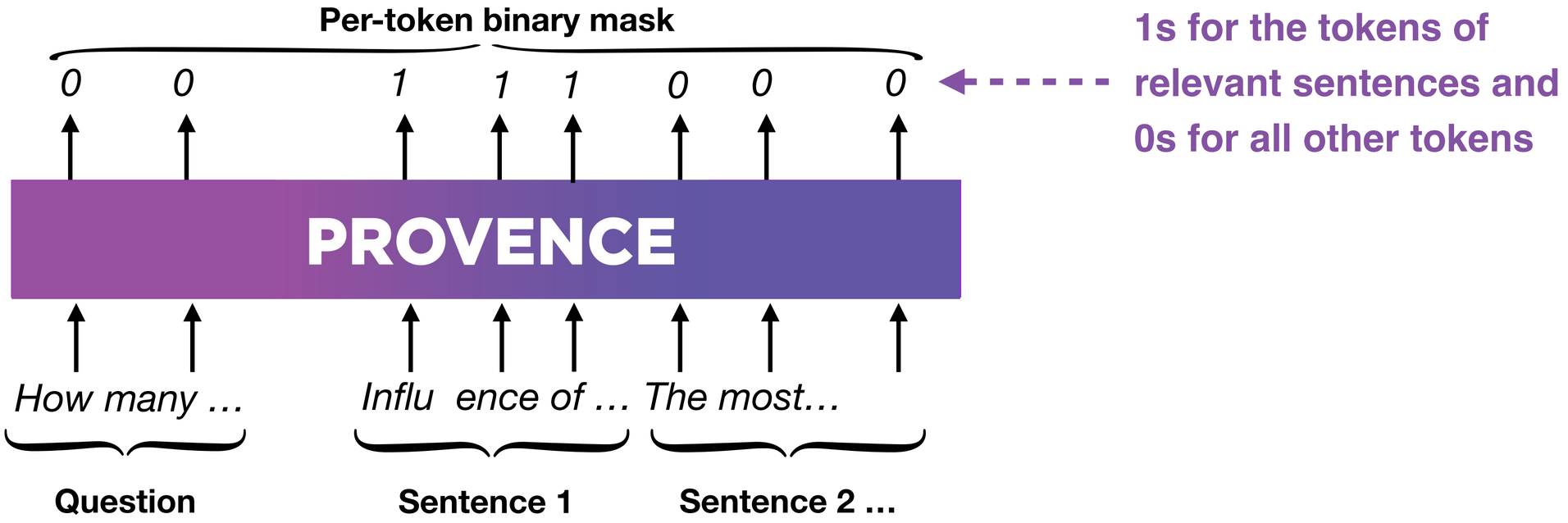
Influence of French on English. The most notable influence of French on English has been its extensive contribution to the English lexicon. It has been estimated that about a third of the words in English are French in origin. Linguist Henriette Walter claims that this total may be as high as two thirds. Linguist Anthony Lacoudre has estimated that over 40,000 English words come directly from French.

How many French words are there in English?

Train question

Step 3: training Provence

We tune a DeBERTa model on a *sequence labeling task*:



Useful properties of Provence architecture (1)

- Provence encodes all the sentences and the question **together**,
  to better understand which sentences to keep

- In contrast, prior works process sentences **independently one-by-one**,
  losing their context and making errors in context pruning

Question: Can you eat pumpkin every day?

Retrieved context: Pumpkin is at its peak in the fall. Eating pumpkin every day can help reduce inflammation, strengthen your immune system and promote eye health. It may also help lower blood pressure. However, if you eat too much, you may experience diarrhea from a high dose of fiber. Read on to learn all about pumpkin's nutrition and health benefits.

These sentences are unclear without the preceding sentences, i.e. that it is about pumpkin

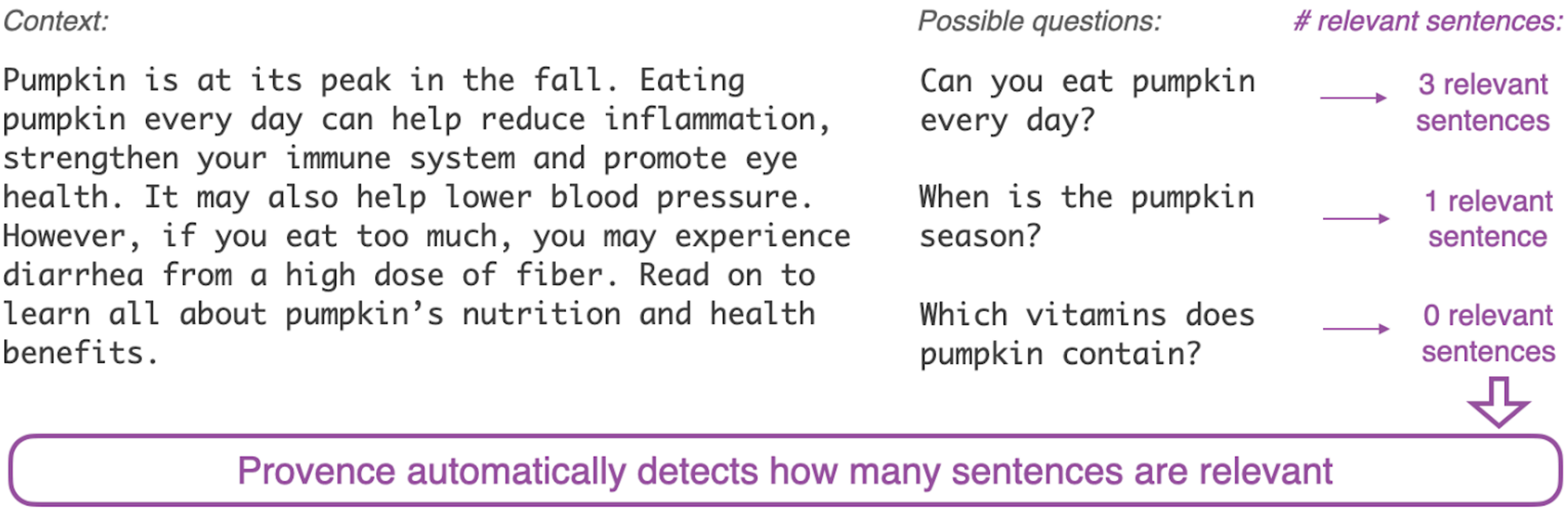
⇩

Provence encodes all sentences and the question together, to better understand which sentences to keep

Useful properties of Provence architecture (2)

- Provence dynamically determines **how many sentences to keep** for each question-context pair

- In contrast, prior works select **a fixed number of sentences**



Useful properties of Provence architecture (3)

- Provence is **efficient**, due to the extractive task formulation and compact base model

- In contrast, prior approaches often rely on **billion-size models** or generate the pruned context **autoregressively**

Useful properties of Provence architecture (3)

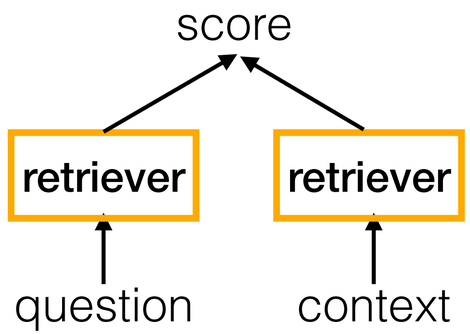
- Provence is **efficient**, due to the extractive task formulation and compact base model

- In contrast, prior approaches often rely on **billion-size models** or generate the pruned context **autoregressively**

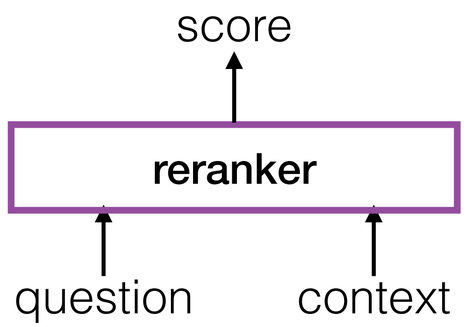
Let’s make Provence
even more efficient,
i.e. almost zero-cost!

A more detailed RAG pipeline

- fast, but less precise first stage of retrieval
- slower, but more precise second stage of retrieval



at inference, only need to encode a query → fast (contexts are pre-encoded)

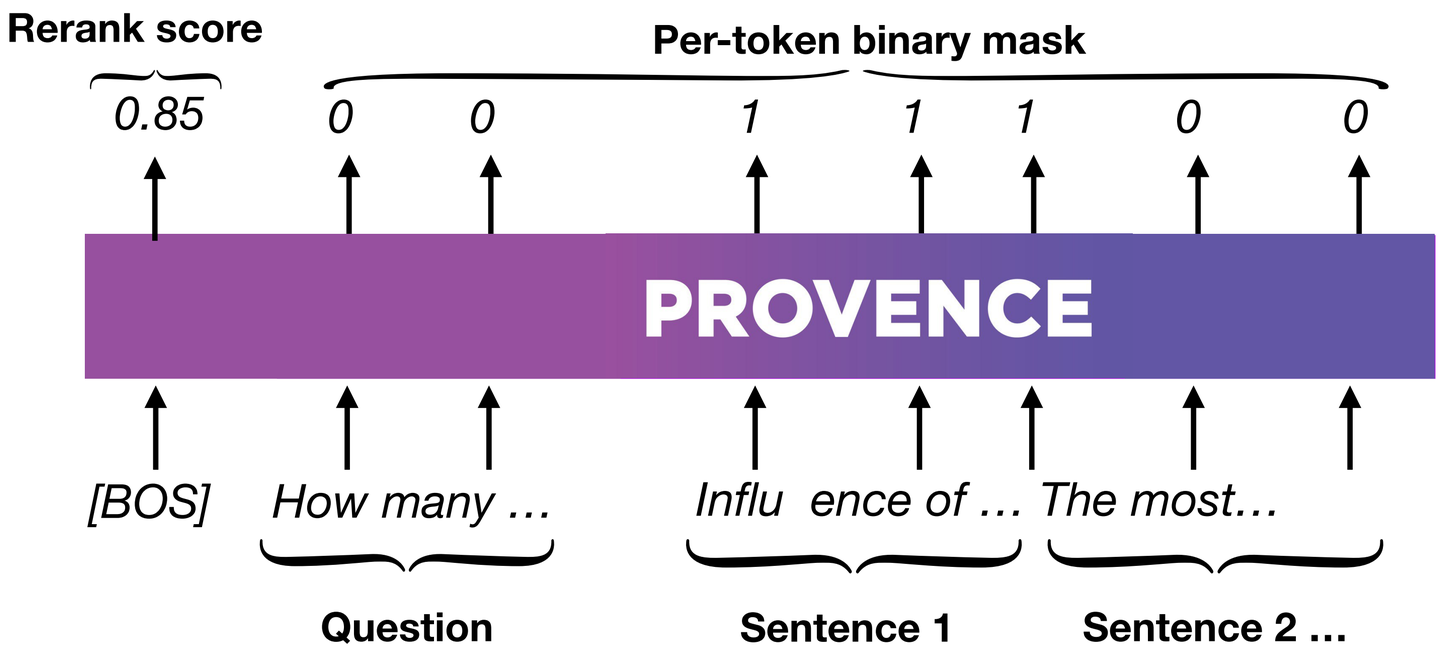


at inference, need to encode each context with a query → slower

Same architecture as Provence!

We propose to **augment a reranker,** an already existing part of the RAG pipeline, **with context pruning capabilities!**

Context pruning & reranking in a single model

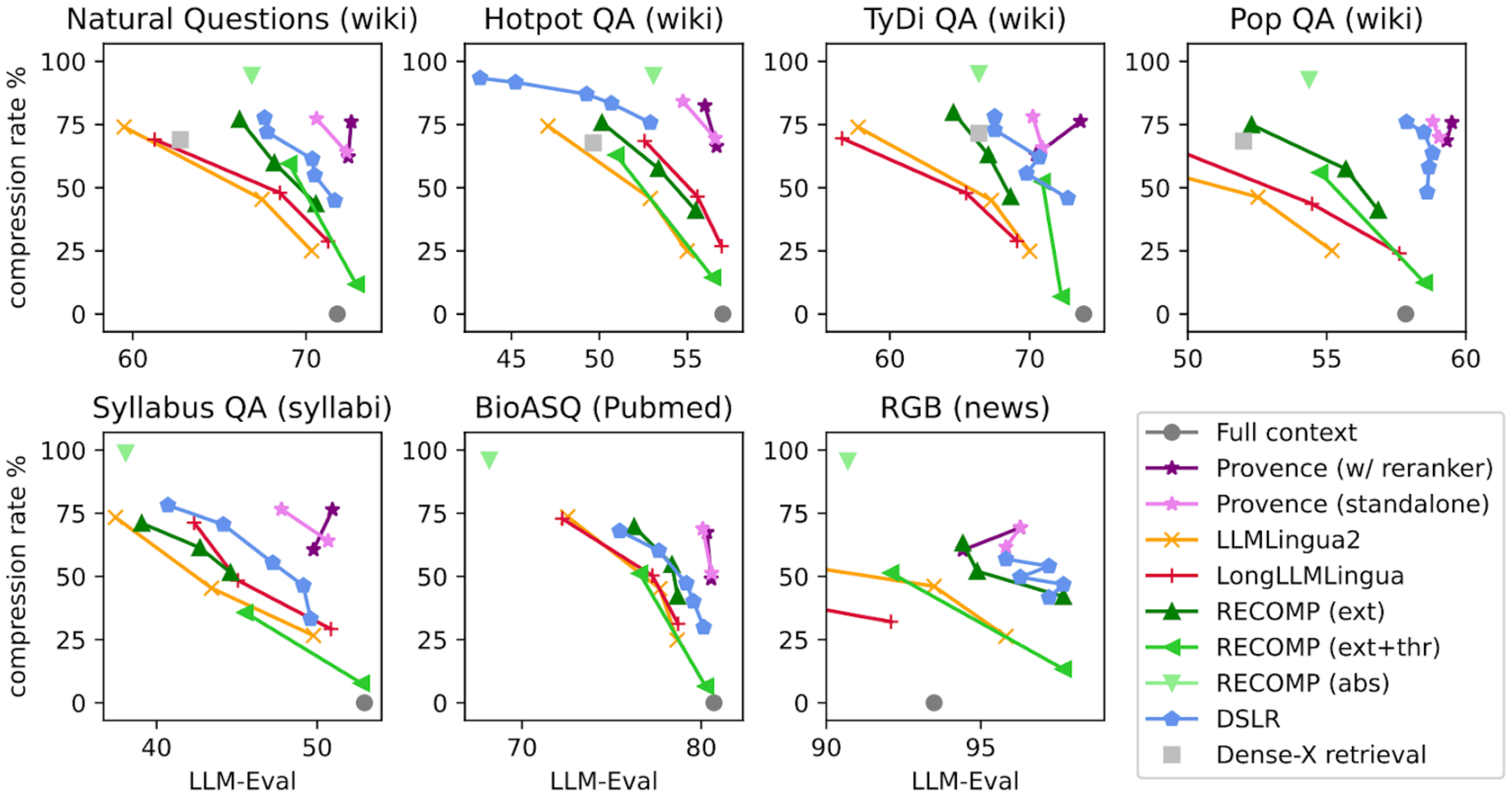


- In addition to the reranking head, we add a second prediction head for context pruning
- Start from a pretrained reranker and tune it with a two-task objective
- Augmenting reranker with context pruning does not hurt reranking performance!

Since we are reusing the forward pass of reranking, **context pruning comes at almost zero cost**!

Main results

top right corner corresponds to the best model



Provence consistently outperforms other approaches, in all domains, and stays on the Pareto front.

Provence is the only model that performs context pruning with little-to-no drop in performance

PROVENCE Pruning and Reranking Of retrieVEd relevaNt ContExts

- An approach for training an **adaptable**, **robust**, and **efficient** (zero-cost!) context pruner, ready to be used **out-of-the-box** for any QA domain and any LLM

- Key ingredients of the approach:
  - formulating context pruning as **sequence labeling**
  - unifying **context pruning and reranking in a single model**
  - training on **diverse data**

- Provence enables **state-of-the-art context pruning**, with little-to-no performance drop across various domains

Model: https://huggingface.co/naver/provence-reranker-debertav3-v1

Blog: https://huggingface.co/blog/nadiinchi/provence

Code: https://github.com/naver/bergen/tree/main/scripts/provence