Paper Review

ATLAS
: Few-shot Learning with Retrieval Augmented Language Models

Gautier Izacard, Patrick Lewis et al.

Meta AI, 2022

Yejin Yoon

NLP Lab.

Dept. of Artificial Intelligence Application

Hanyang University

stillwithyou@hanyang.ac.kr

What are Covered in this Presentation

- **Details of Atlas's method**
  - **Atlas :** *Izacard*, *Gautier*, et al. "Few-shot learning with retrieval augmented language models." arXiv preprint arXiv:2208.03299 (2022).

- **Some Details of predecessors**
  - **Contriever :** *Izacard*, *Gautier*, et al. "Towards unsupervised dense information retrieval with contrastive learning." arXiv preprint arXiv:2112.09118 (2021).
  - **FiD :** *Izacard*, *Gautier*, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." arXiv preprint arXiv:2007.01282 (2020).
  - **FiD-KD** : *Izacard*, *Gautier*, and Edouard Grave. "Distilling knowledge from reader to retriever for question answering." arXiv preprint arXiv:2012.04584 (2020).

- **Some Details of Related papers**
  - **MoCo :** He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

What are NOT Covered in this Presentation

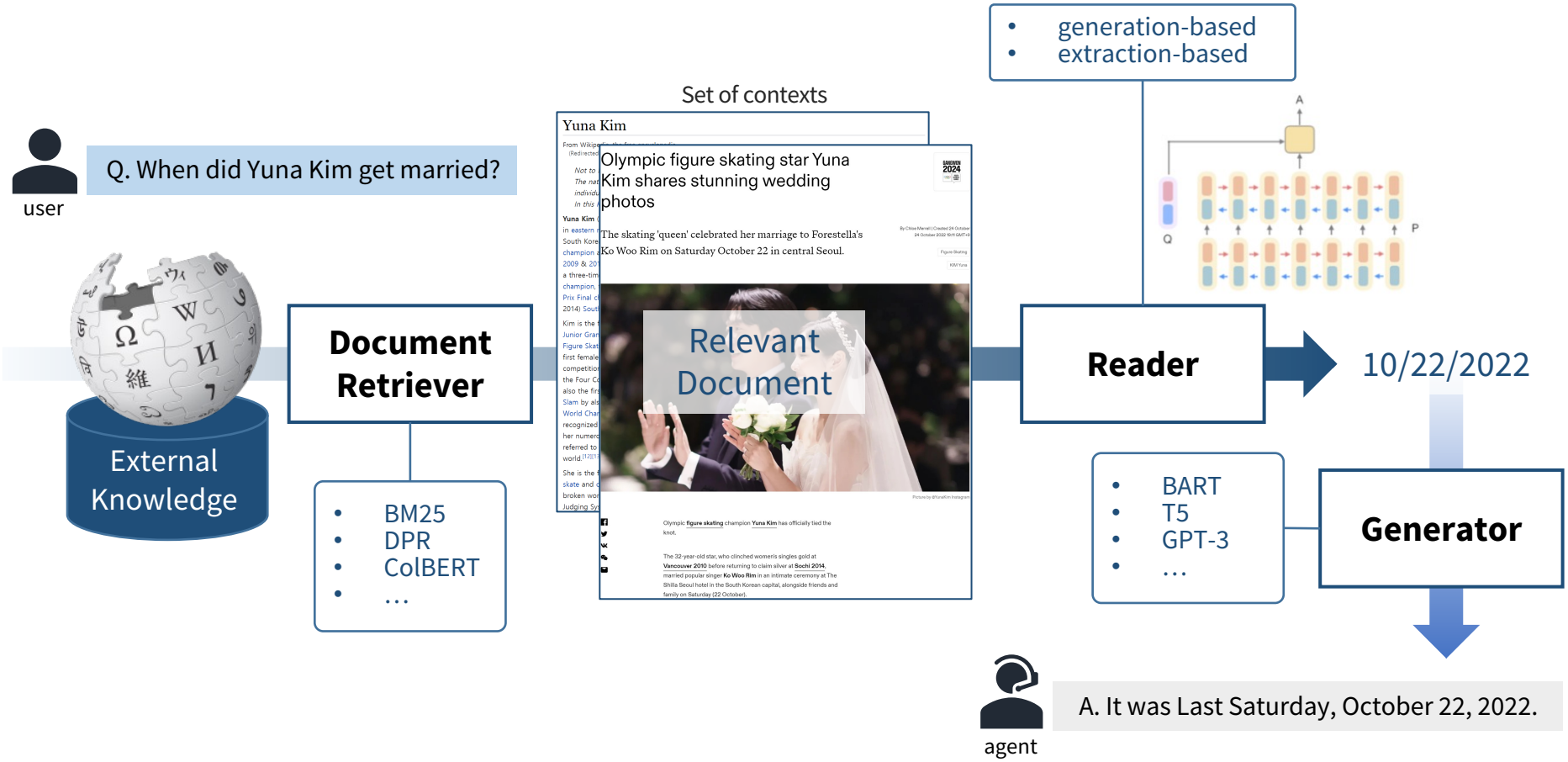
- **Details of predecessors from authors**
  - **Contriever :** *Izacard Gautier*, et al. "Towards unsupervised dense information retrieval with contrastive learning." arXiv preprint arXiv:2112.09118 (2021).
  - **FiD :** *Izacard Gautier*, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." arXiv preprint arXiv:2007.01282 (2020).
  - **FiD-KD** : *Izacard Gautier*, and Edouard Grave. "Distilling knowledge from reader to retriever for question answering." arXiv preprint arXiv:2012.04584 (2020).
  - **RAG** : *Lewis Patrick*, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
  - **KILT** : Petroni, Fabio, et al. "KILT: a benchmark for knowledge intensive language tasks." arXiv preprint arXiv:2009.02252 (2020).

1. Pre-Requisites

- What is **ODQA**? (Review)
- What is **Semi-parametric method**?

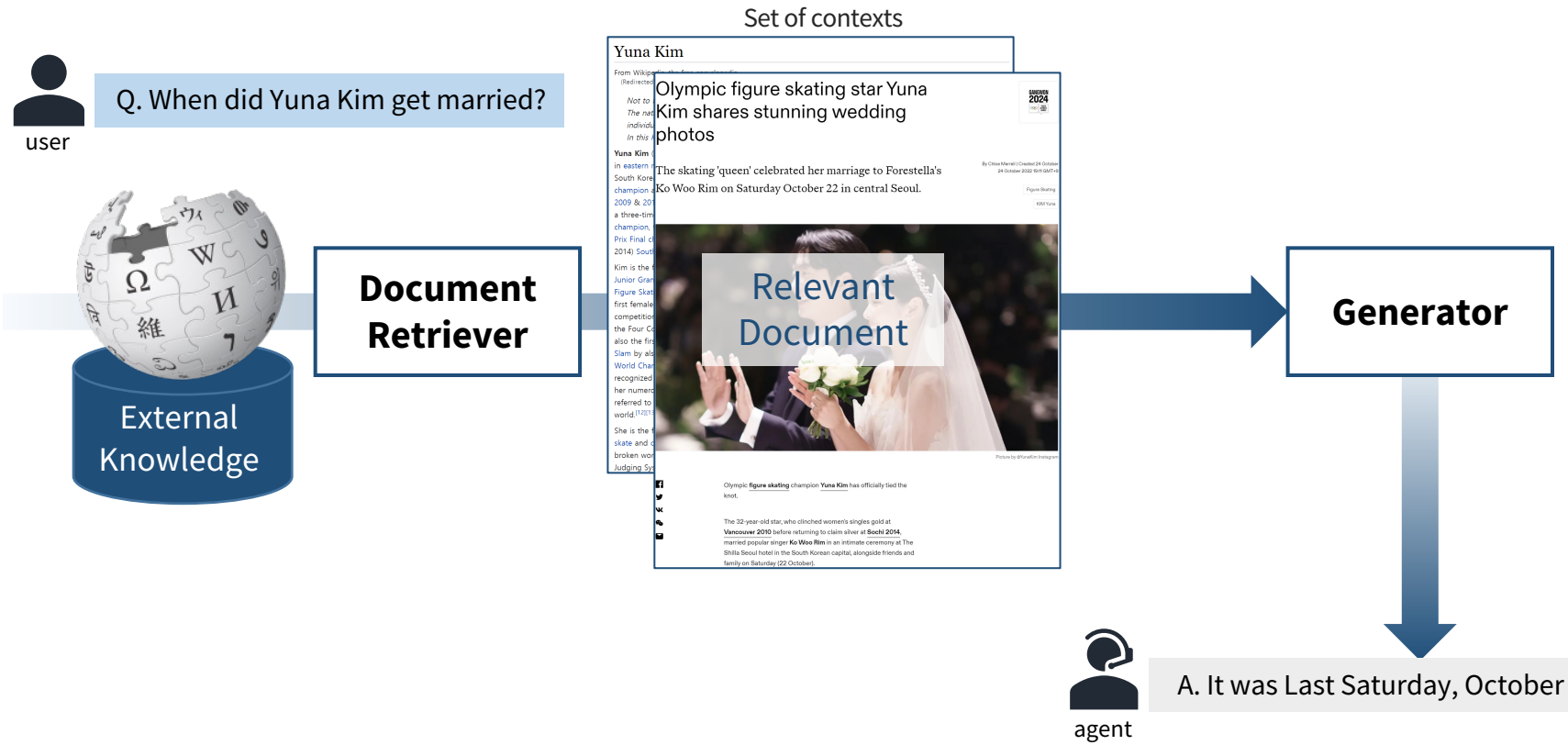
Pre-Requisites : What is ODQA?

- **Pipeline approaches (3-stage retriever-reader-generator system)**



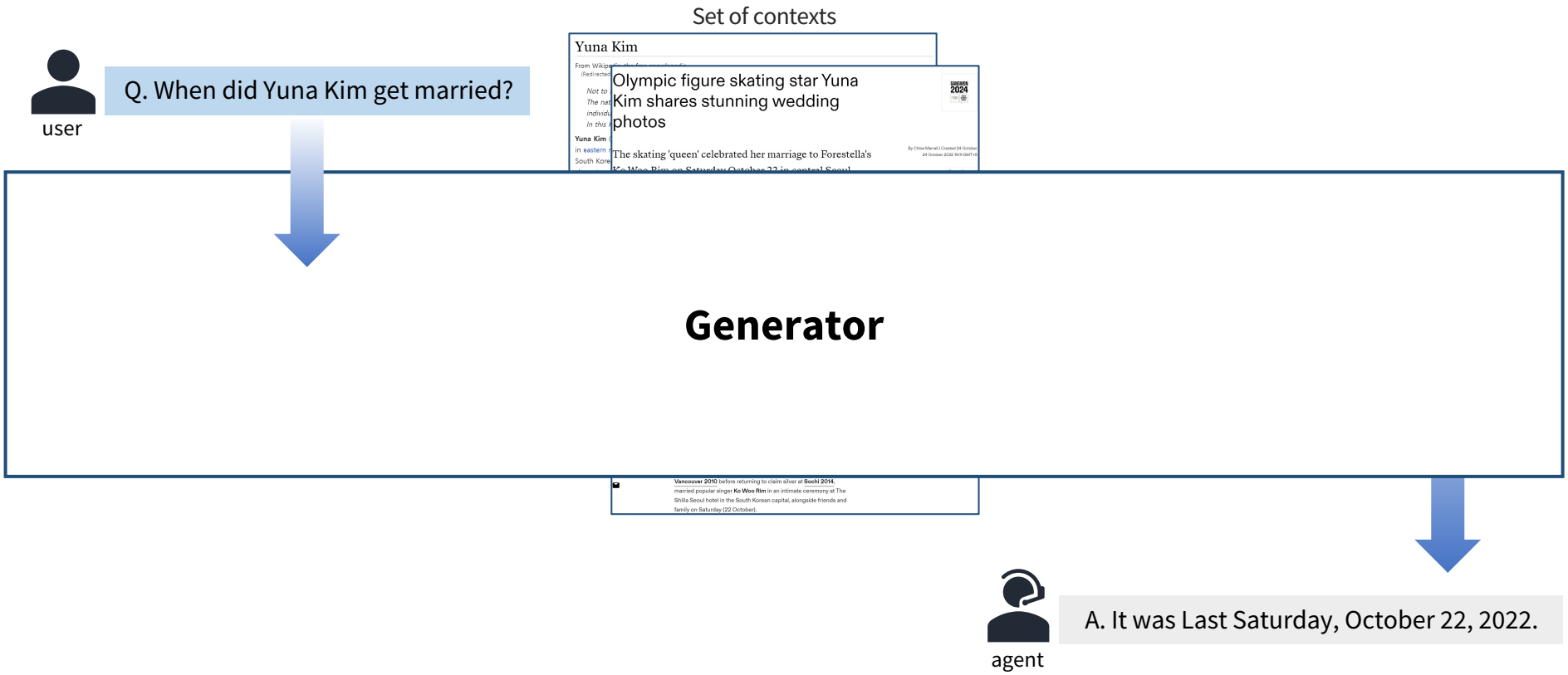
Pre-Requisites : What is ODQA?

- **Pipeline approaches (2-stage retriever-generator system)**



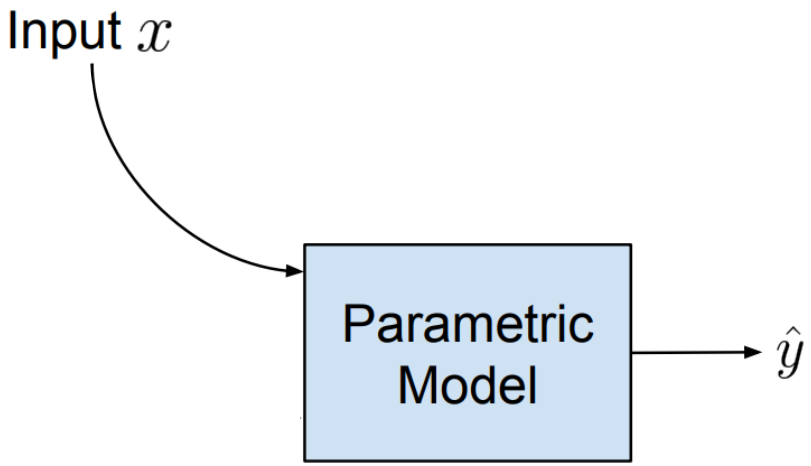
Pre-Requisites : What is ODQA?

- **End-to-end approaches: Closed-book QA**



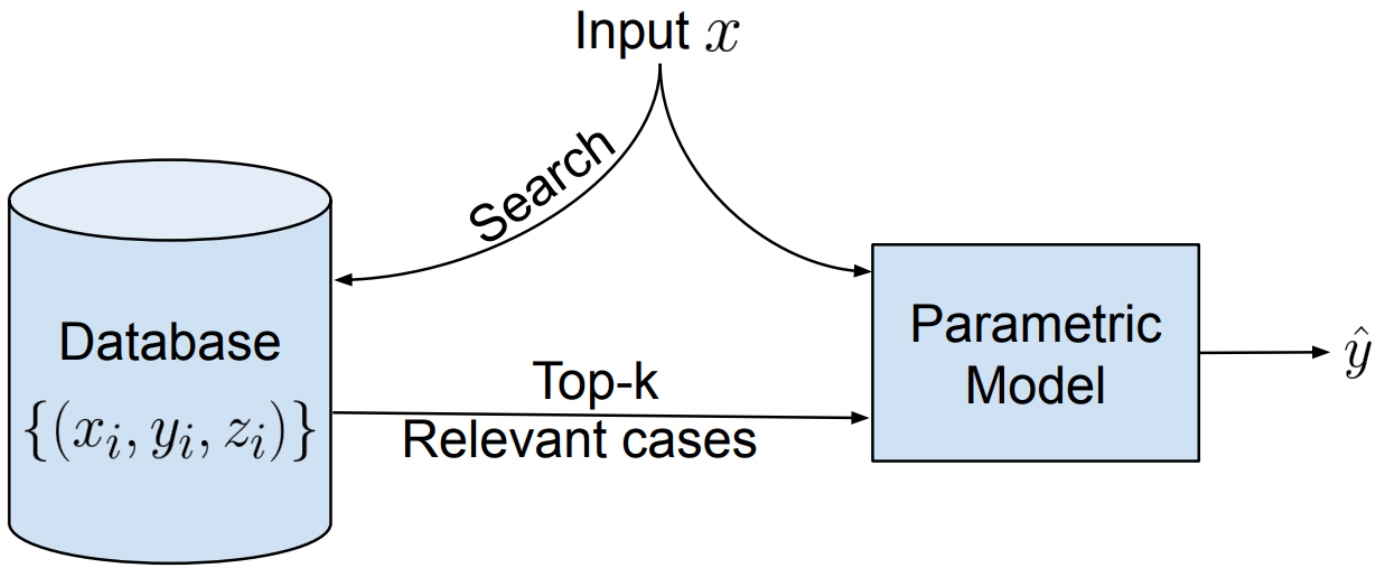
Pre-Requisites : What is Semi-parametric?

- **Semi-parametric = parametric method + non-parametric method**
  Fit with knowledge-intensive task



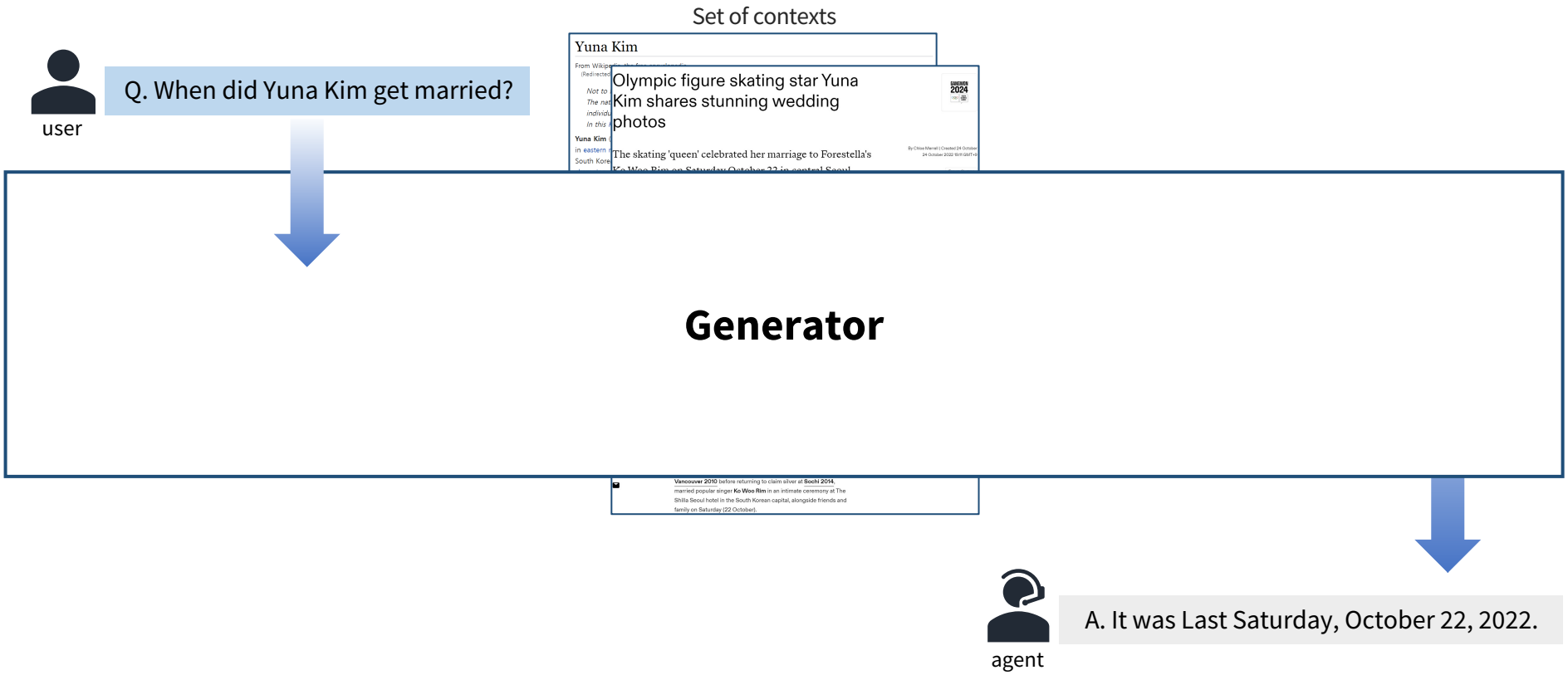
Pre-Requisites : What is Semi-parametric?

- **Semi-parametric = parametric method + non-parametric method**
  Fit with knowledge-intensive task



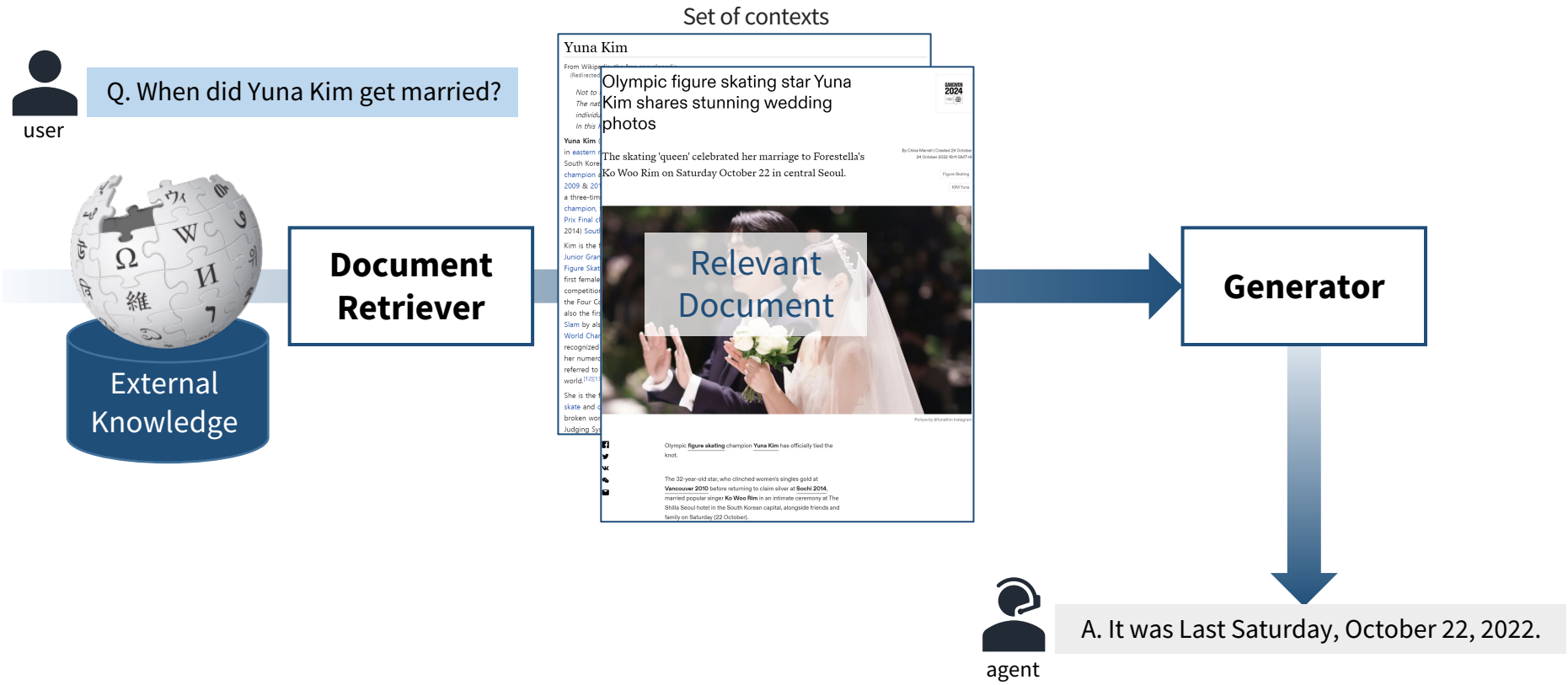
Pre-Requisites : What is Semi-parametric?

- **Parametric method**



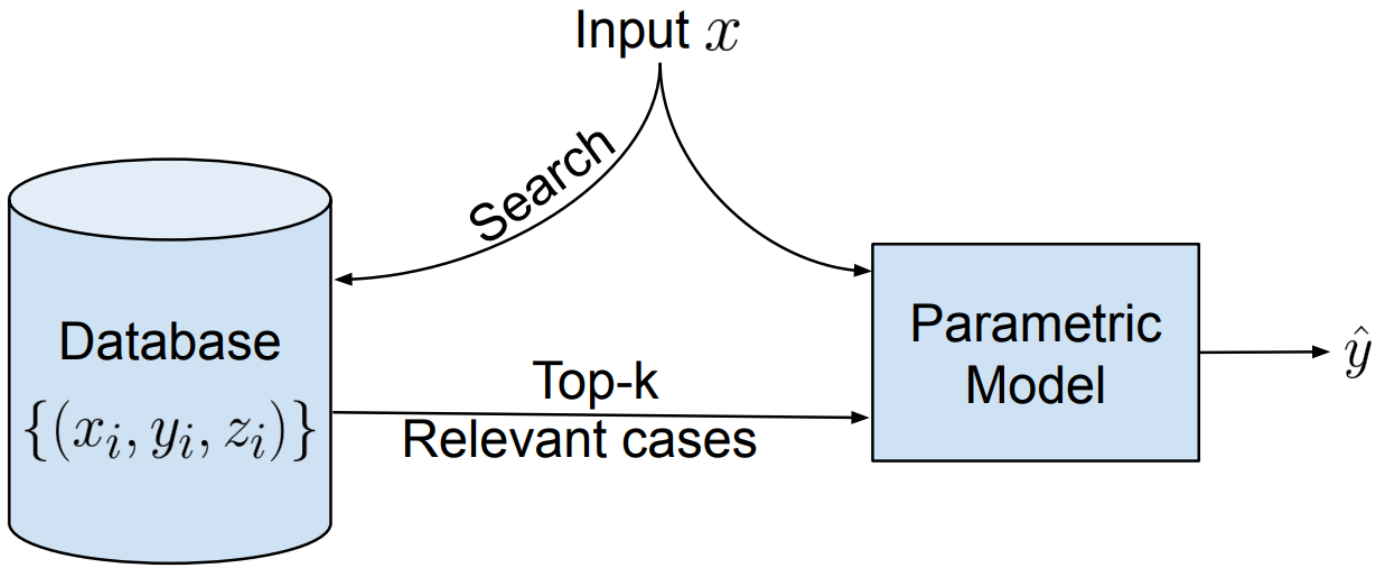
Pre-Requisites : What is Semi-parametric?

- **Semi-parametric method**



Pre-Requisites : What is Semi-parametric?

- **Semi-parametric = parametric method + non-parametric method**
  Fit with knowledge-intensive task



- Interpretable & controllable
  - Control DB → Control $\hat{y}$
- Easy to update DB
- Fixed LM

ATLAS
: Few-shot Learning with Retrieval Augmented Language Models

Gautier Izacard, Patrick Lewis et al.

Meta AI, 2022

Yejin Yoon

Contents

2. Introduction

- TL; DR
- Background
- Problem States

TL; DR

- **ATLAS is retrieval-augmented LM capable of strong few-shot learning, despite having lower parameter #.**
  - 2-phrase semi-parametric architecture:

    Contriever *(Izacard et al., 2022)* + **FiD** (Izacard & Grave, 2020)

    Retriever — Language Model
  - Experiments on different training strategies for each module

Background

- **Limitation of LLM**
  - LLM have shown impressive few-shot results on a wide range of tasks.
  - LLM also owe this implement to both a larger computational budget and the ability to memorize more information.
  - LLM do not track the cause of their predictions because they do not have an explicit causal model or a transparent representation of their internal workings.

- **IR system fit with knowledge-intensive tasks**
  - Employing a retrieval-augmented architecture, memory can be outsourced and replaced by an external non-parametric knowledge source.
  - This non-parametric knowledge source can enhance a parametric language model.
  - In addition, such architectures are attractive due to a number of other established advantages in terms of adaptability, interpretability and efficiency

Problem States

- **How to design and train retrieval-augmented language models, with a focus on downstream few-shot learning and sample efficiency**
  - Exploration of fine-tuning strategies to efficiently adapt both the retriever and the language model to the task at hand

3. Main Points: Method

- Architecture
- Training objectives for retriever
- Pretext task
- Fine-tuning

Methods

- **Text-to-text framework (T5)**

  : the system gets *a text query* as input, and generates *a text output*

Task	Query	Output
Fact Checking	*Bermuda Triangle is in the western part of the Himalayas.*	*False*
Question Answering	*who is playing the halftime show at super bowl 2016*	*Coldplay*
Entity Linking	*NTFS-3G is an open source <E>cross-platform</E> implementation of the Microsoft Windows NTFS file system with read-write support.*	*Cross-platform software*

Figure 2: Examples of query and output pairs for different tasks from KILT.

Methods

- **Text-to-text framework (T5)**

  : the system gets *a text query* as input, and generates *a text output*

Task	Query	Output
Fact Checking	*Bermuda Triangle is in the western part of the Himalayas.*	*False*
Question Answering	*who is playing the halftime show at super bowl 2016*	*Coldplay*
Entity Linking	*NTFS-3G is an open source <E>cross-platform</E> implementation of the Microsoft Windows NTFS file system with read-write support.*	*Cross-platform software*

Figure 2: Examples of query and output pairs for different tasks from KILT.

- **KILT: a Benchmark for Knowledge Intensive Language Tasks** (2021, Meta AI)
  - a set of benchmarks for evaluating language models that have been infused with external knowledge
  - benchmark consists of a series of tasks that test the ability of models to use knowledge to perform various NLP tasks; ODQA, Dialogue, Fact checking
    - to measure the performance of models in terms of accuracy, efficiency, and interpretability
    - to compare different models and to evaluate their strengths and weaknesses in using knowledge to improve their performance

Methods

- **Text-to-text framework (T5)**

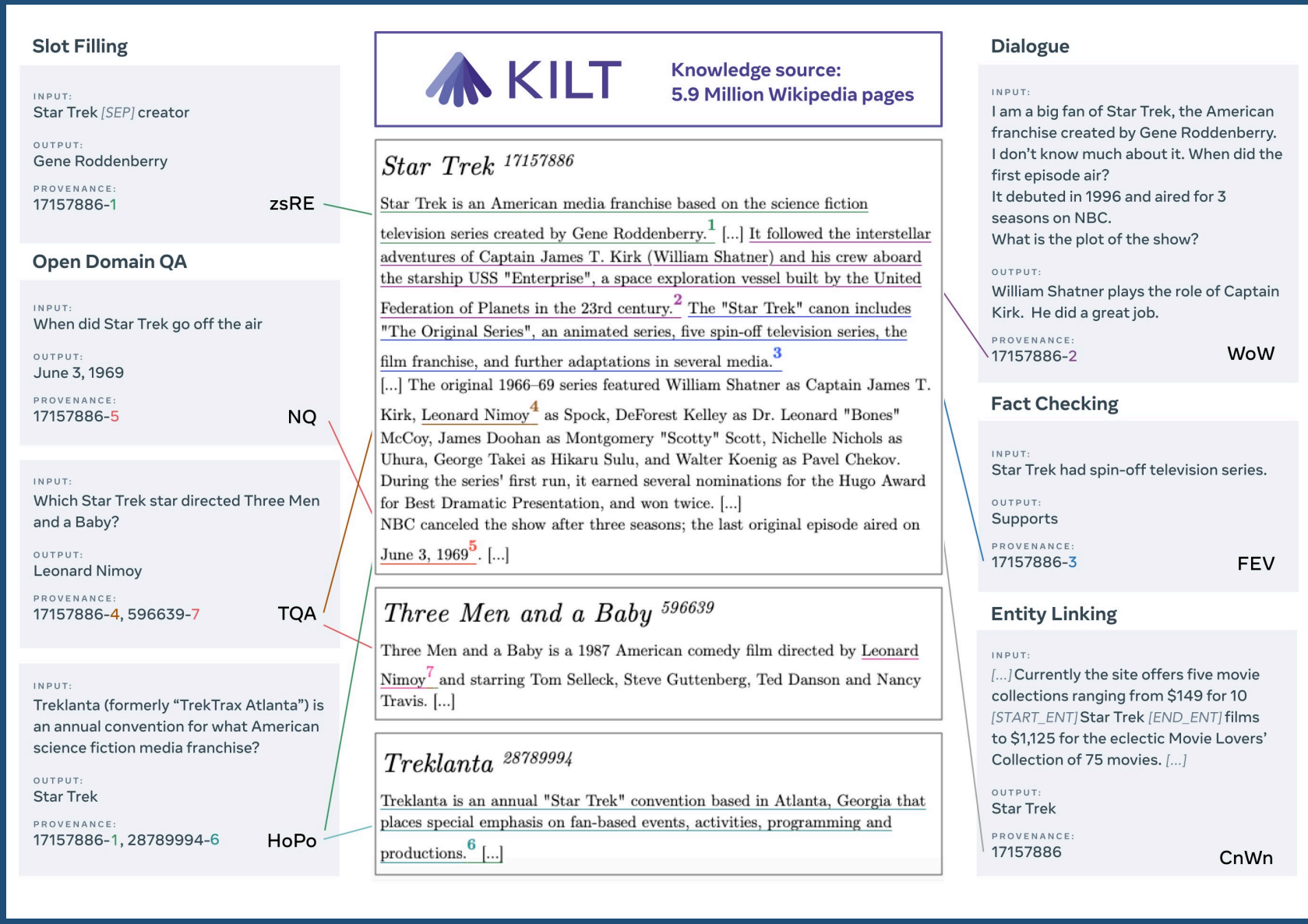
  : the system gets *a text query* as input, and generates *a text output*

Task	Query	Output
Fact Checking	*Bermuda Triangle is in the western part of the Himalayas.*	*False*
Question Answering	*who is playing the halftime show at super bowl 2016*	*Coldplay*
Entity Linking	*NTFS-3G is an open source <E>cross-platform</E> implementation of the Microsoft Windows NTFS file system with read-write support.*	*Cross-platform software*

Figure 2: Examples of query and output pairs for different tasks from KILT.

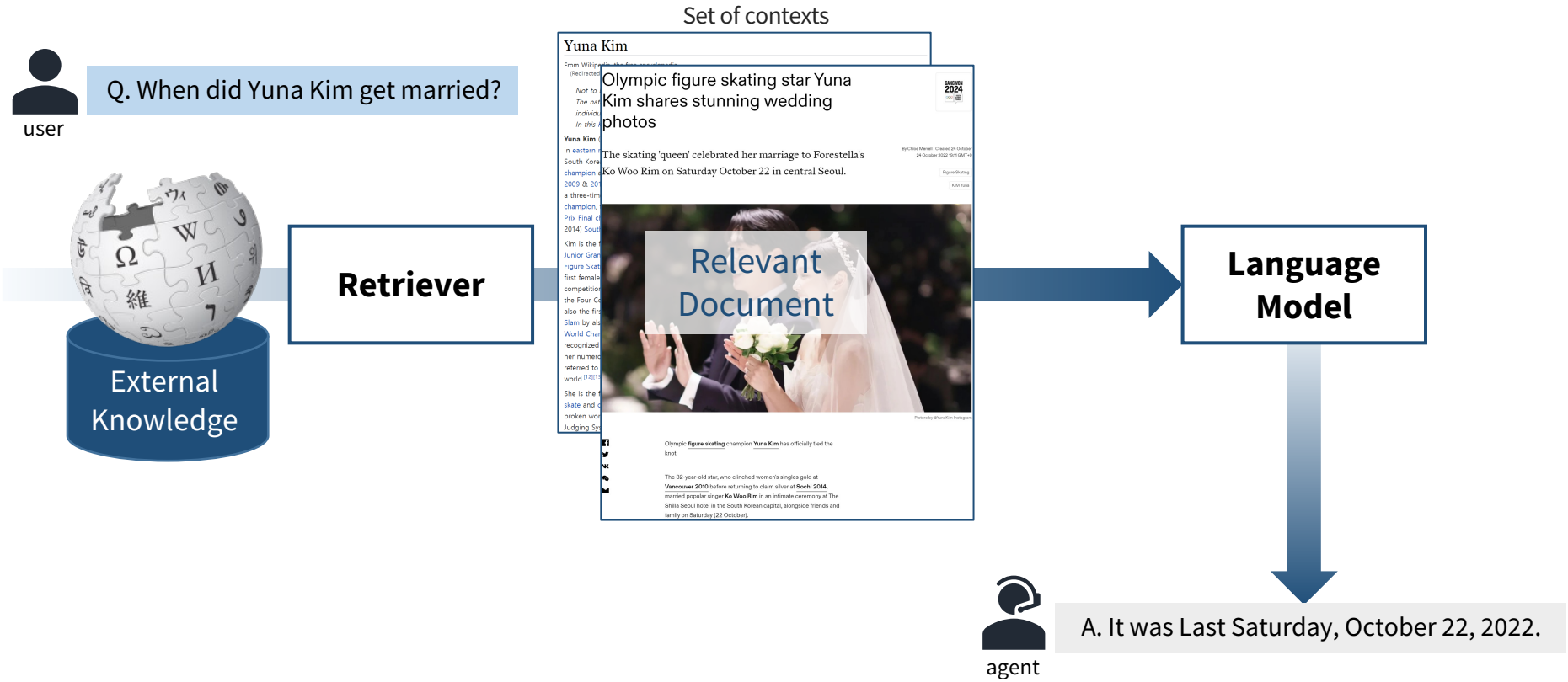
Label	Dataset	Reference	Task	Input Format	Output Format
FEV	FEVER	Thorne et al. (2018a)	Fact Checking	Claim	Classification
AY2	AIDA CoNLL-YAGO	Hoffart et al. (2011b)	Entity Linking	Text Chunk	Entity
WnWi	WNED-WIKI	Guo and Barbosa (2018)	Entity Linking	Text Chunk	Entity
WnCw	WNED-CWEB	Guo and Barbosa (2018)	Entity Linking	Text Chunk	Entity
T-REx	T-REx	Elsahar et al. (2018)	Slot Filling	Structured	Entity
zsRE	Zero Shot RE	Levy et al. (2017)	Slot Filling	Structured	Entity
NQ	Natural Questions	Kwiatkowski et al. (2019)	Open Domain QA	Question	Extractive
HoPo	HotpotQA	Yang et al. (2018)	Open Domain QA	Question	Short Abstractive
TQA	TriviaQA	Joshi et al. (2017)	Open Domain QA	Question	Extractive
ELI5	ELI5	Fan et al. (2019b)	Open Domain QA	Question	Long Abstractive
WoW	Wizard of Wikipedia	Dinan et al. (2019)	Dialogue	Conversation	Long Abstractive

Methods


Slot Filling
INPUT:
Star Trek [SEP] creator
OUTPUT:
Gene Roddenberry
PROVENANCE:
17157886-1
zsRE
Open Domain QA
INPUT:
When did Star Trek go off the air
OUTPUT:
June 3, 1969
PROVENANCE:
17157886-5
NQ
INPUT:
Which Star Trek star directed Three Men and a Baby?
OUTPUT:
Leonard Nimoy
PROVENANCE:
17157886-4, 596639-7
TQA
INPUT:
Treklanta (formerly "TrekTrax Atlanta") is an annual convention for what American science fiction media franchise?
OUTPUT:
Star Trek
PROVENANCE:
17157886-1, 28789994-6
HoPo
KILT
Knowledge source:
5.9 Million Wikipedia pages
Star Trek 17157886
Star Trek is an American media franchise based on the science fiction television series created by Gene Roddenberry.[1] [...] It followed the interstellar adventures of Captain James T. Kirk (William Shatner) and his crew aboard the starship USS "Enterprise", a space exploration vessel built by the United Federation of Planets in the 23rd century.[2] The "Star Trek" canon includes "The Original Series", an animated series, five spin-off television series, the film franchise, and further adaptations in several media.[3]
[...] The original 1966–69 series featured William Shatner as Captain James T. Kirk, Leonard Nimoy[4] as Spock, DeForest Kelley as Dr. Leonard "Bones" McCoy, James Doohan as Montgomery "Scotty" Scott, Nichelle Nichols as Uhura, George Takei as Hikaru Sulu, and Walter Koenig as Pavel Chekov. During the series' first run, it earned several nominations for the Hugo Award for Best Dramatic Presentation, and won twice. [...]
NBC canceled the show after three seasons; the last original episode aired on June 3, 1969[5]. [...]
Three Men and a Baby 596639
Three Men and a Baby is a 1987 American comedy film directed by Leonard Nimoy[7] and starring Tom Selleck, Steve Guttenberg, Ted Danson and Nancy Travis. [...]
Treklanta 28789994
Treklanta is an annual "Star Trek" convention based in Atlanta, Georgia that places special emphasis on fan-based events, activities, programming and productions.[6] [...]
Dialogue
INPUT:
I am a big fan of Star Trek, the American franchise created by Gene Roddenberry. I don't know much about it. When did the first episode air?
It debuted in 1996 and aired for 3 seasons on NBC.
What is the plot of the show?
OUTPUT:
William Shatner plays the role of Captain Kirk. He did a great job.
PROVENANCE:
17157886-2
WoW
Fact Checking
INPUT:
Star Trek had spin-off television series.
OUTPUT:
Supports
PROVENANCE:
17157886-3
FEV
Entity Linking
INPUT:
[...] Currently the site offers five movie collections ranging from $149 for 10 [START_ENT] Star Trek [END_ENT] films to $1,125 for the eclectic Movie Lovers' Collection of 75 movies. [...]
OUTPUT:
Star Trek
PROVENANCE:
17157886
CnWn

Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - retrieve the top-k relevant documents from a large corpus of text
  - Feed these documents to the language model, along with the query



Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - retrieve the top-k relevant documents from a large corpus of text
  - Feed these documents to the language model, along with the query

Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - A dual-encoder architecture: the query and documents are embedded independently by a transformer encoder
  - Average pooling is applied over the outputs of the last layer to obtain one vector representation per query or document
  - A similarity score btw the query and each document: the dot product between their corresponding embeddings
  - Pre-training Loss: MoCo contrastive Loss (He et al. 2020)
  - Pre-training Data: Unsupervised data only → train w/o document annotation

Contrastive Learning

Methods: Architecture

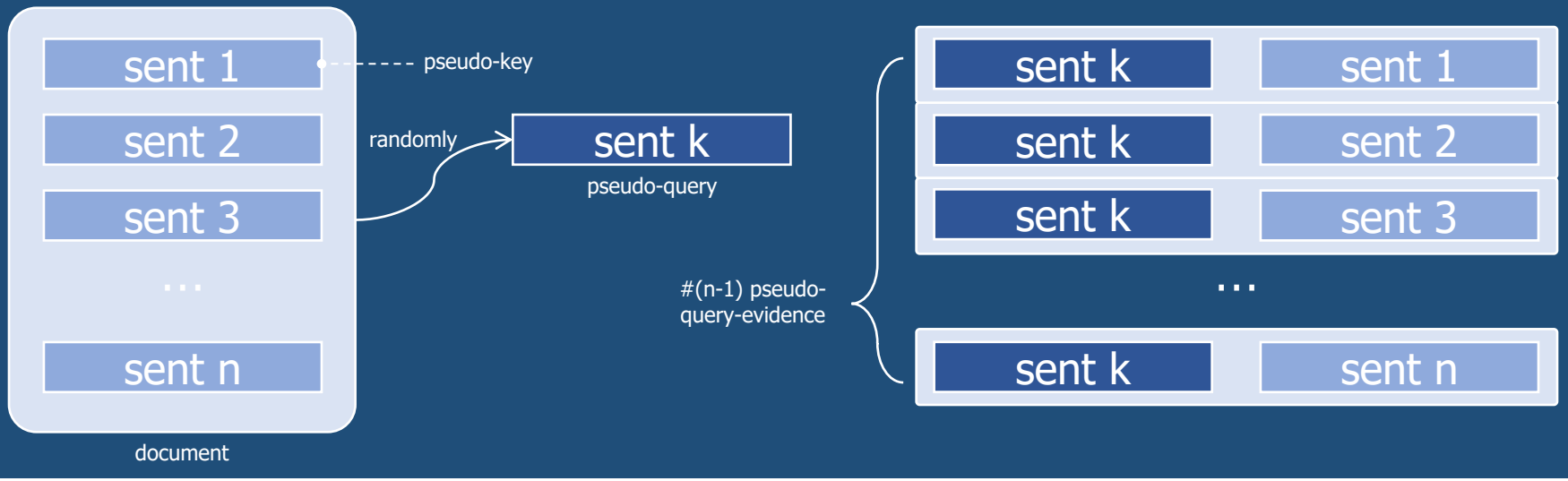
- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)

Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

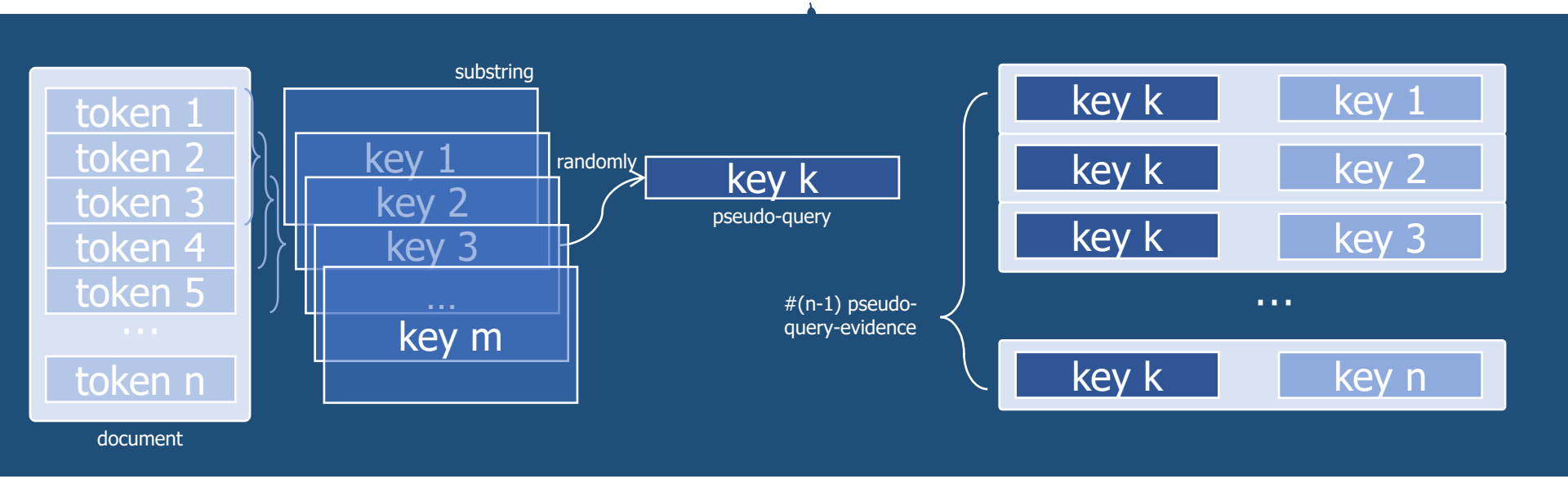
- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)



Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

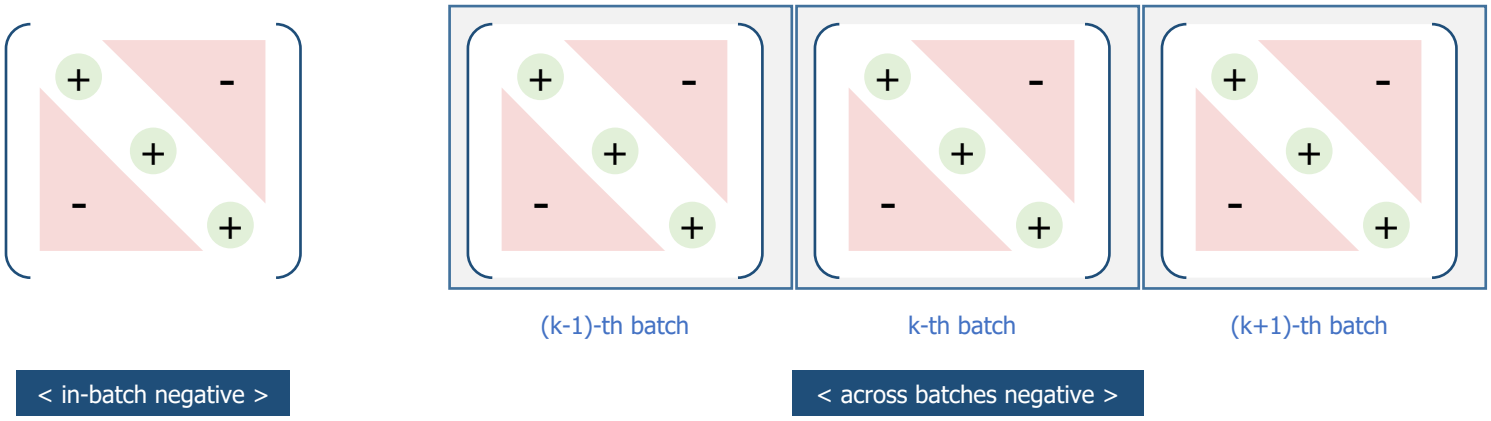
- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)



Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

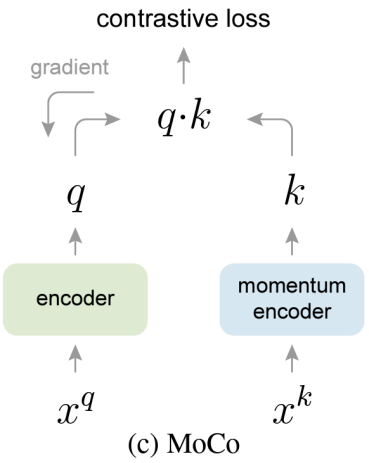
- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)
  - Negative pair: Negative pairs across batches methods



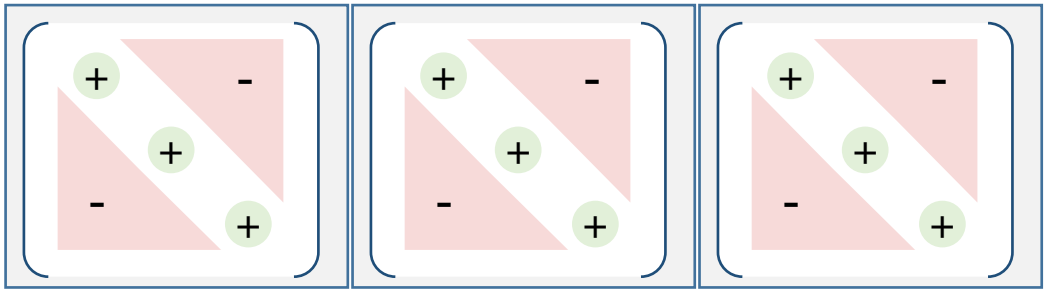
Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)
  - Negative pair: Negative pairs across batches methods



(c) MoCo



(k-1)-th batch k-th batch (k+1)-th batch

< across batches negative >

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$: Momentum Contrastive

Methods: Architecture

- **2 sub-models: the retriever + the language model**
  - Retriever retrieves the top-k relevant documents from a large corpus of text
  - These documents are fed to the language model, along with the query

- **Contriever (Izacard et al., 2022)**: an information retrieval technique based on continuous dense embeddings
  - Positive pair: Independent Cropping based on ICT(Inverse Close Task)
  - Negative pair: Negative pairs across batches methods

- **Fusion-in-Decoder (Izacard & Grave, 2020)**:

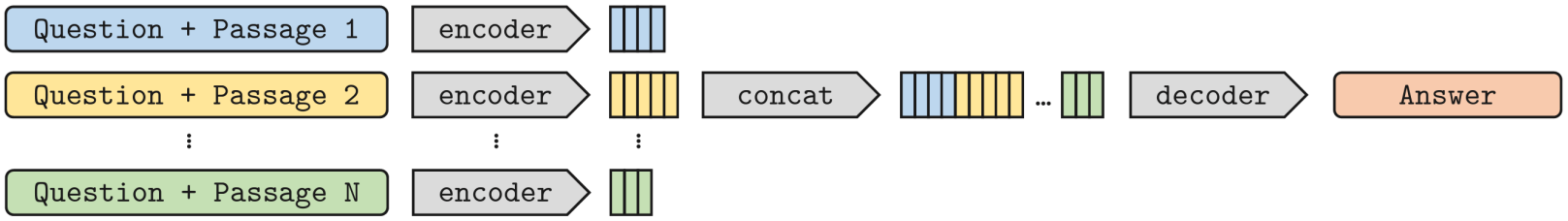
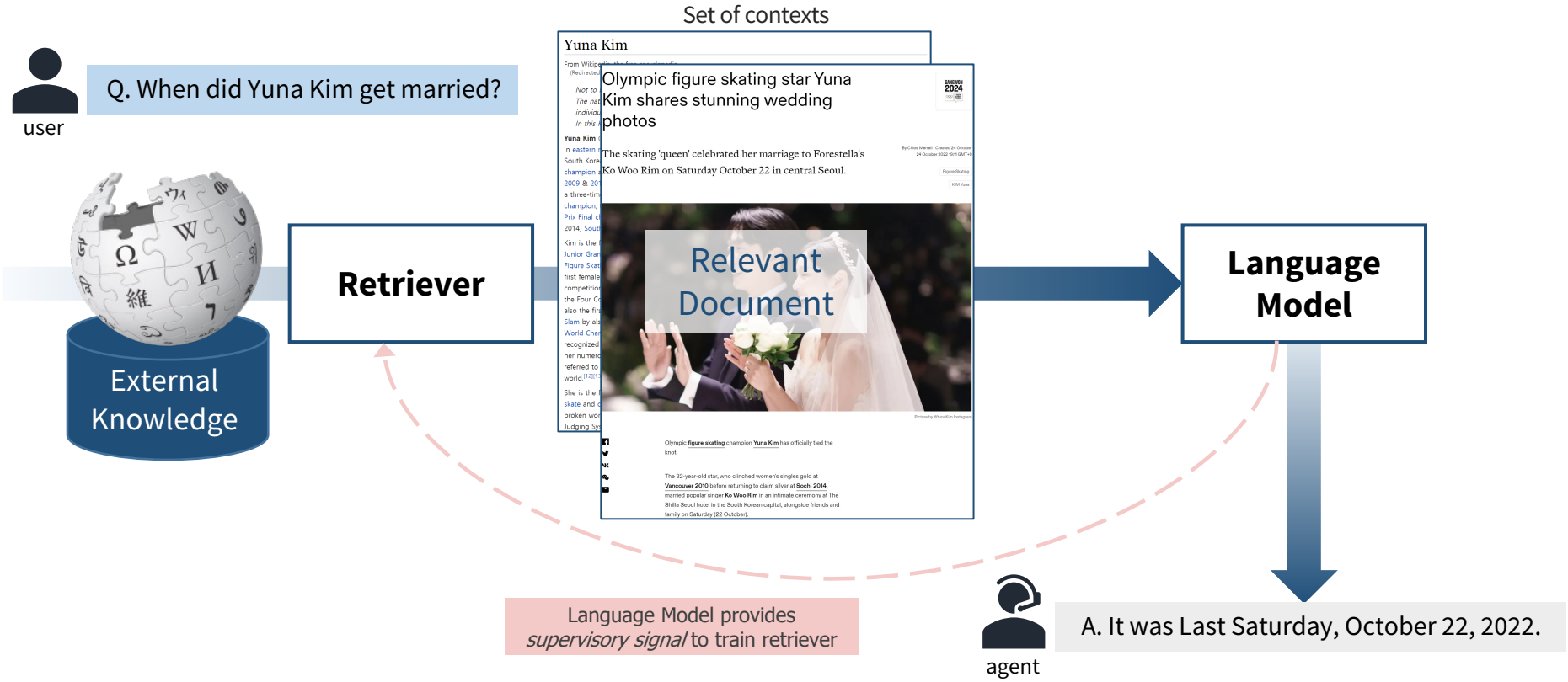


Figure 2: Architecture of the Fusion-in-Decoder method.

Methods: Training Objectives for the Retriever



- **If the language model finds a document useful when generating the output, the retriever objective should encourage the retriever to rank said document higher.**

Methods: Training Objectives for the Retriever

1. Attention Distillation (A Dist)

$$\mathrm{KL}(p_{ATTN} \,||\, p_{RETR}) = \sum_{k=1}^{K} p_{ATTN}(\mathbf{d}_k) \log\left(\frac{p_{ATTN}(\mathbf{d}_k)}{p_{RETR}(\mathbf{d}_k)}\right)$$

2. End-to-end training of Multi-Document Reader and Retriever ($EMDR^2$)

$$\log\left[\sum_{k=1}^{K} p_{LM}(\mathbf{a} \mid \mathbf{q}, \mathbf{d}_k)\, p_{RETR}(\mathbf{d}_k \mid \mathbf{q})\right]$$

3. Perplexity Distillation (P Dist)

$$\mathrm{KL}(p_k \,||\, p_{\mathrm{RETR}}) = \sum_{k=1}^{K} p_k(\mathbf{d}_k) \log\left(\frac{p_k(\mathbf{d}_k)}{p_{\mathrm{RETR}}(\mathbf{d}_k)}\right)$$

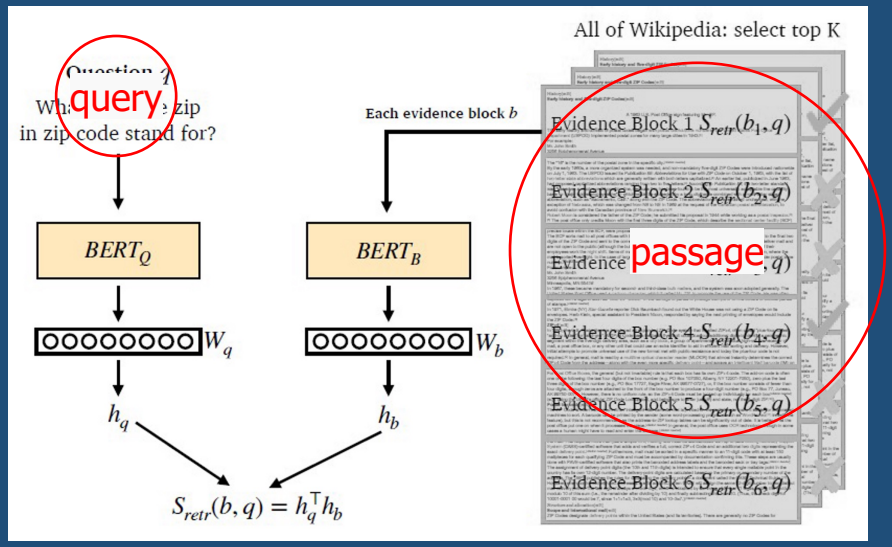
4. Leave-one-out Perplexity Distillation (LOOP)

$$\mathrm{KL}(p_{\mathrm{LOOP}} \,||\, p_{\mathrm{RETR}}) = \sum_{k=1}^{K} p_{\mathrm{LOOP}}(\mathbf{d}_k) \log\left(\frac{p_{\mathrm{LOOP}}(\mathbf{d}_k)}{p_{\mathrm{RETR}}(\mathbf{d}_k)}\right)$$

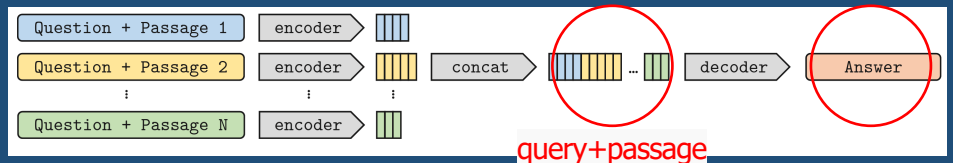
Methods: Training Objectives for the Retriever

1. Attention Distillation (A Dist)

$$\mathrm{KL}(p_{ATTN} \,||\, p_{RETR}) = \sum_{k=1}^{K} p_{ATTN}(\mathbf{d}_k) \log\left(\frac{p_{ATTN}(\mathbf{d}_k)}{p_{RETR}(\mathbf{d}_k)}\right)$$



< previous pre-training >



< jointly pre-training >

Methods: Training Objectives for the Retriever

1. Attention Distillation (A Dist)

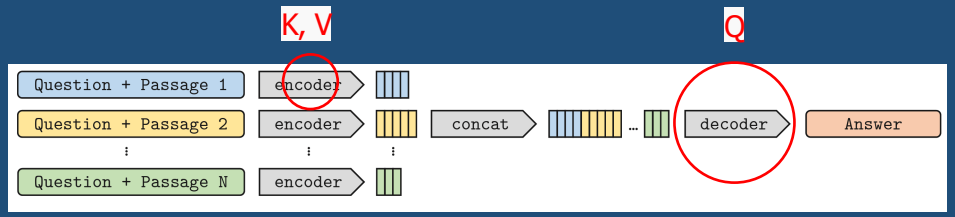
$$\mathrm{KL}(p_{ATTN} \,||\, p_{RETR}) = \sum_{k=1}^{K} p_{ATTN}(\mathbf{d}_k) \log\left(\frac{p_{ATTN}(\mathbf{d}_k)}{p_{RETR}(\mathbf{d}_k)}\right)$$

$$p_{RETR}(\mathbf{d} \mid \mathbf{q}) = \frac{\exp\left(\frac{s(\mathbf{d},\mathbf{q})}{\theta}\right)}{\sum_{k=1}^{K} \exp\left(\frac{s(\mathbf{d}_k,\mathbf{q})}{\theta}\right)}$$

* $s(\mathbf{d},\mathbf{q})$: dot-product btw the query and documents vectors
* θ: a temperature hyper-parameter

p_{ATTN}: a slight variation of FiD-KD decoder's cross-attention score

$$\tilde{G}_{q,p} = \frac{\exp(G_{q,p})}{\sum_{p' \in \mathcal{D}_q} \exp(G_{q,p'})}$$



Methods: Training Objectives for the Retriever

1. Attention Distillation (A Dist)

$$\mathrm{KL}(p_{ATTN} \,||\, p_{RETR}) = \sum_{k=1}^{K} p_{ATTN}(\mathbf{d}_k) \log\left(\frac{p_{ATTN}(\mathbf{d}_k)}{p_{RETR}(\mathbf{d}_k)}\right)$$

2. End-to-end training of Multi-Document Reader and Retriever ($EMDR^2$)

$$\log\left[\sum_{k=1}^{K} p_{LM}(\mathbf{a} \mid \mathbf{q}, \mathbf{d}_k)\, p_{RETR}(\mathbf{d}_k \mid \mathbf{q})\right]$$

3. Perplexity Distillation (P Dist)

$$p_k = \frac{\exp(\log p_{LM}(\mathbf{a} \mid \mathbf{d}_k, \mathbf{q}))}{\sum_{k=1}^{K} \exp(\log p_{LM}(\mathbf{a} \mid \mathbf{d}_i, \mathbf{q}))}$$

$$\mathrm{KL}(p_k \,||\, p_{\mathrm{RETR}}) = \sum_{k=1}^{K} p_k(\mathbf{d}_k) \log\left(\frac{p_k(\mathbf{d}_k)}{p_{\mathrm{RETR}}(\mathbf{d}_k)}\right)$$

4. Leave-one-out Perplexity Distillation (LOOP)

$$p_{\mathrm{LOOP}} = \frac{\exp(-\log p_{LM}(\mathbf{a} \mid \mathcal{D}_K \setminus \{\mathbf{d}_k\}, \mathbf{q}))}{\sum_{k=1}^{K} \exp(\log p_{LM}(\mathbf{a} \mid \mathcal{D}_K \ \{\mathbf{d}_i,\}\, \mathbf{q}))}$$

$$\mathrm{KL}(p_{\mathrm{LOOP}} \,||\, p_{\mathrm{RETR}}) = \sum_{k=1}^{K} p_{\mathrm{LOOP}}(\mathbf{d}_k) \log\left(\frac{p_{\mathrm{LOOP}}(\mathbf{d}_k)}{p_{\mathrm{RETR}}(\mathbf{d}_k)}\right)$$

Methods: Pretext task

1. **Prefix language modeling**
2. **Masked language modeling (T5)**
3. **Title to section generation**

Methods: Fine-tuning

- **Efficient Retriever fine-tuning**
  - Offline indexing(ColBERT)
- **Full index update**

$$\frac{N \times P_{RETR}}{4 \times B \times K \times P_{LM} \times R}, \qquad \frac{P_{RETR}}{P_{LM}} \approx \frac{1}{25}, \qquad \Rightarrow \frac{N}{100 \times B \times K \times R}$$

  - N: # of documents
  - B: batch size
  - R: training step
  - K: # of retrieved document
  - P_{RETR}: # of params of retriever
  - P_{LM}: # of params of LM
- **Query-side fine-tuning**
  - Update only Query encoder (Document encoder is frozen)
- **Re-ranking** $\quad \frac{L \times P_{RETR}}{4 \times K \times P_{LM}}$

4. Effects: Experiments

- Benchmarks
- Experiment setting
- Evaluation

Effects: Benchmarks

- **KILT**
  - a set of benchmarks for evaluating language models that have been infused with external knowledge
- **MMLU: Massively-Multitask Language Understanding**
  - 57 multi-choice question answering datasets
  - High school mathematics, professional law, logical fallacies, etc.
  - Focused on few-shot setting
- **Original Question Answering datasets**
  - NQ, TriviaQA, FEVER
  - TempLAMA

Effects: Details of experiment setting

- **Pre-training**
  - unsupervised contriever (pre-trained T5-lm-adapt)
- **Fine-tuning**
  - Fixed every hyper-parameter
  - 50-iterations (64-shot), 200-iterations (1024-shot)
  - 32-batch, 4×10^{-5} learning rate w/ linear decay, 5 warmup step
- **Unlabeled datasets**
  - For retriever training
  - Wikipedia dump(2021.12.20), Common Crawl dump(2020.10)
  - Document Filtering is similar to Gopher's method (350M passages)

Effects: Evaluation Results

- **RQ1) Does jointly pre-training the whole model lead to better few-shot performance?**

- **RQ2) What is the best objective function for the retriever, and the best pretext task?**

Table 1: **Retriever loss ablation.** We compare different loss functions to pre-train the retriever jointly with the language model. We use the prefix MLM task, and the December 2021 Wikipedia dump for both the index and pre-training data. Fine-tuning is performed with query-side fine-tuning and the loss used for pre-training. Best result is bold, second highest underlined.

		64-shot				1024-shot			
	MLM	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
No Joint pre-training	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9
Fixed retriever	0.823	39.9	14.3	72.4	42.2	45.3	17.9	90.0	51.1
ADist	0.780	40.9	14.4	73.8	43.0	46.2	17.2	**90.9**	**51.4**
EMDR^2	0.783	43.3	14.6	72.1	43.3	44.9	**18.3**	85.7	49.6
PDist	0.783	**45.0**	**15.0**	**77.0**	**45.7**	44.9	17.9	90.2	51.0
LOOP	**0.766**	41.8	**15.0**	74.4	43.7	**47.1**	17.9	87.5	50.8

Effects: Evaluation Results

- **RQ2) What is the best objective function for the retriever, and the best pretext task?**

Table 2: **Pretext task ablation.** We compare different pretext tasks, used to jointly pre-train our models. Examples are randomly sampled from the training set of the KILT version of the dataset. We report the exact match on NaturalQuestions, the F1 score on Wizard of Wikipedia and the accuracy on FEVER.

	64-shot				1024-shot			
	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Prefix Language Modelling	41.0	14.5	64.9	40.1	**44.7**	17.9	86.0	49.5
Masked Language Modelling	**42.7**	**14.9**	**69.7**	**42.4**	**44.7**	**18.3**	**88.8**	**50.6**
Title-to-section generation	41.1	15.2	66.1	40.8	45.4	17.9	84.6	49.3

Table 3: **Index content ablation.** In this table, we report results for models where the content of the index was changed between the pre-training and the fine-tuning.

		64-shot				1024-shot			
Index	Training data	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Wiki	Wiki	**42.7**	14.9	69.7	**42.4**	44.7	18.3	88.8	**50.6**
Wiki	CC	40.9	**15.3**	67.3	41.2	**44.8**	**18.4**	88.1	50.4
CC	Wiki	32.9	14.5	**72.1**	39.8	37.8	17.1	85.8	46.9
CC	CC	38.4	14.9	70.1	41.1	42.0	17.3	**88.9**	49.4

Effects: Evaluation Results

- **RQ3) How to efficiently fine-tune Atlas on tasks with limited training data?**

Table 4: **Retriever fine-tuning ablation.** Here, we compare different strategies to fine-tune the retriever in a few-shot setting.

	64-shot				1024-shot			
	NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Standard fine-tuning	44.3	14.9	73.2	44.1	47.0	18.4	89.7	**51.7**
Top-100 re-ranking	44.2	14.6	75.4	**44.7**	**47.1**	**18.7**	88.9	51.6
Query-side fine-tuning	**45.0**	**15.0**	**77.0**	**45.7**	44.9	17.9	**90.2**	51.0
Fixed retriever	36.8	14.5	72.0	41.1	38.0	17.7	89.3	48.3

Effects: Evaluation Results

- **MMLU Result**

Table 7: **Comparison to state-of-the-art on MMLU.** *For the 5-shot setting, ATLAS uses fine-tuning, while previous works use in-context learning. The ATLAS model uses de-biased inference. Train FLOPS refers to total the amount of computation necessary to train the model, including pre-training and/or fine-tuning.

Setting	Model	Params	Train FLOPS	All	Hum.	Soc. Sci.	STEM	Other
zero-shot	ATLAS	11B	3.5e22	47.1	43.6	54.1	38.0	54.4
5-shot	GPT-3	175B	3.1e23	43.9	40.8	50.4	36.7	48.8
	Gopher	280B	5.0e23	60.0	56.2	71.9	47.4	66.1
	Chinchilla	70B	5.0e23	**67.5**	**63.6**	**79.3**	**55.0**	73.9
	ATLAS*	11B	3.5e22	47.9	46.1	54.6	38.8	52.8
5-shot (multi-task)	ATLAS	11B	3.5e22	56.6	50.1	66.4	46.4	66.2
Full / Transfer	UnifiedQA	11B	3.3e22	48.9	45.6	56.6	40.2	54.6
	GPT-3	175B	3.1e23	53.9	52.5	63.9	41.4	57.9
	ATLAS	11B	3.5e22	66.0	61.1	77.2	53.2	**74.4**

Effects: Evaluation Results

- **ODQA Result**

Table 8: **Comparison to state-of-the-art on question answering.** We report results on NaturalQuestions, and on TriviaQA for both the filtered set, commonly used for open-domain question answering and the unfiltered hidden set for which evaluation is accessible online: https://competitions.codalab.org/competitions/17208. For the 64-shot setting, our model uses fine-tuning, while the other models use prompting.

	NQ		TriviaQA filtered		TriviaQA unfiltered	
Model	64-shot	Full	64-shot	Full	64-shot	Full
GPT-3 (Brown et al., 2020)	29.9	-	-	-	71.2	-
Gopher (Rae et al., 2021)	28.2	-	57.2	-	61.3	-
Chinchilla (Hoffmann et al., 2022)	35.5	-	64.6	-	72.3	-
PaLM (Chowdhery et al., 2022)	39.6	-	-	-	81.4	-
RETRO (Borgeaud et al., 2021)	-	45.5	-	-	-	-
FiD (Izacard & Grave, 2020)	-	51.4	-	67.6	-	80.1
FiD-KD (Izacard & Grave, 2021)	-	54.7	-	73.3	-	-
R2-D2 (Fajcik et al., 2021)	-	55.9	-	69.9	-	-
ATLAS	**42.4**	**60.4**	**74.5**	**79.8**	**84.7**	**89.4**

Effects: Evaluation Results

- **KILT Result**

Table 10: **Downstream results on the KILT hidden test sets** Downstream metrics are accuracy (AIDA CoNLL-YAGO, FEVER, T-REx, zero-shot RE), exact match (Natural Questions, HotpotQA, TriviaQA), or F1 (Wizard of Wikipedia).

Model	AIDA ACC	FEV ACC	T-REx ACC	zsRE ACC	NQ EM	HoPo EM	TQA EM	WoW F1
GENRE (Cao et al., 2021)	89.9	-	-	-	-	-	-	-
Sphere (Piktus et al., 2021)	-	89.0	81.7	74.2	51.6	38.3	72.7	15.5
SEAL (Bevilacqua et al., 2022)	-	89.5	83.6	74.6	53.7	40.5	70.9	18.3
Re2G (Glass et al., 2022)	-	89.6	**87.7**	-	51.7	-	76.3	18.9
FID with RS (Hofstätter et al., 2022)	-	92.2	85.2	**83.7**	61.2	39.1	**84.6**	20.6
ATLAS, 64-shot	66.5	87.1	58.9	74.9	43.6	34.7	76.4	15.5
ATLAS, full train set	**90.6**	**93.5**	85.1	80.8	**61.3**	**50.6**	84.0	**21.6**

5. Contribution & Opinion

Contribution ~~& Opinion~~

- **Atlas is 2-phrase system: retriever + LM**
- **Retriever and Language Model is jointly pre-trained**
- **Various experiments for retrieval-augmented LM**
- **Suggest most effective method in terms of retriever training / pretext task / finetuning**
- **High performance in knowledge-intensive task with few-shot setting**

- ~~**Details of predecessors from authors…**~~

{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.

Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr