

Interpretability for Information Retrieval

Anja Reusch – Postdoctoral Researcher – NLP Lab, Technion



TECHNION



About Me

Dipl.-Inf. @ TU Dresden
in Germany

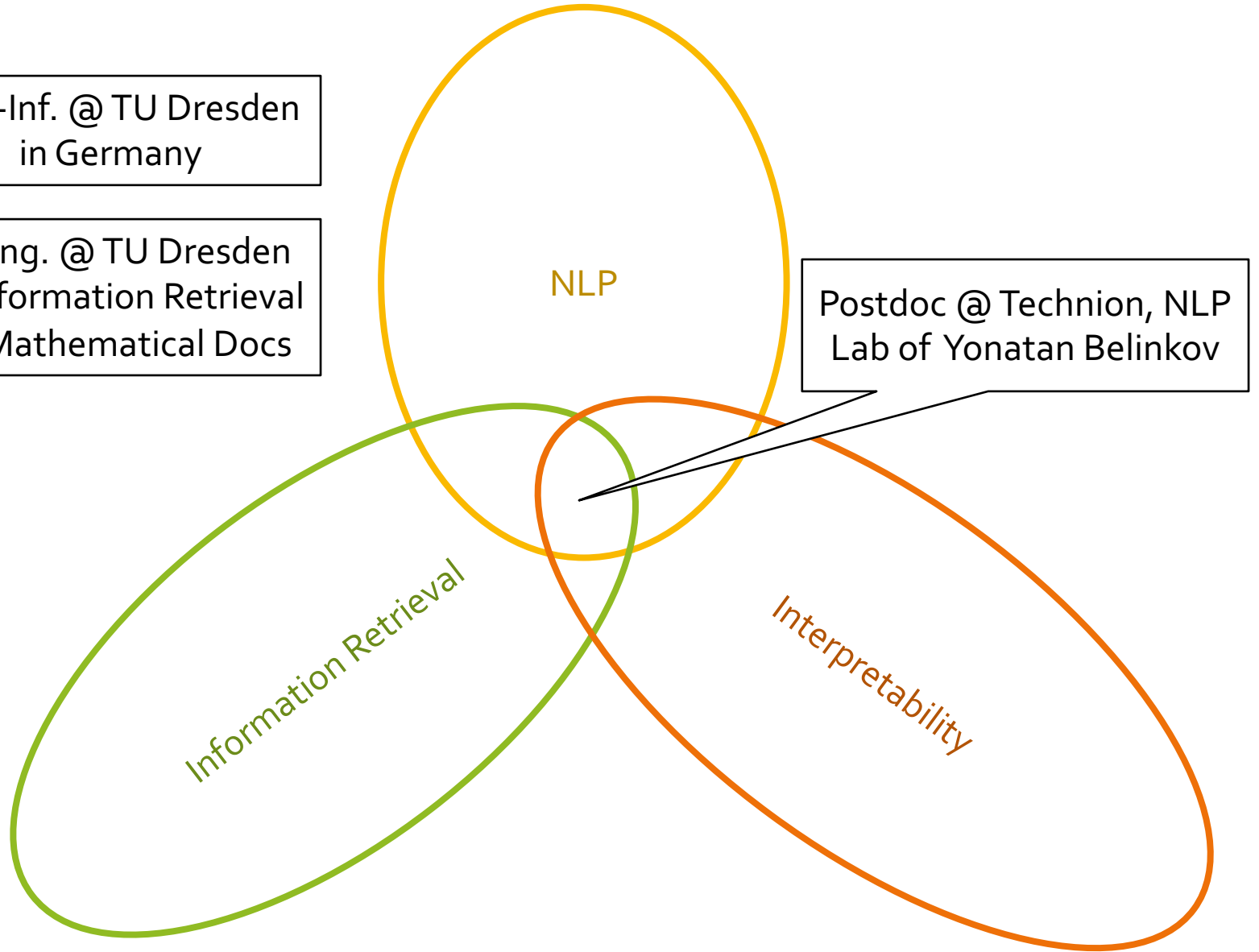
Dr.-Ing. @ TU Dresden
on Information Retrieval
for Mathematical Docs

NLP

Postdoc @ Technion, NLP
Lab of Yonatan Belinkov

Information Retrieval

Interpretability



Interpretability for Retrieval



I want to know
about the latest
developments in
robotics.

Interpretability for Retrieval



I want to know
about the latest
developments in
robotics.



SEARCH



Top Search Results

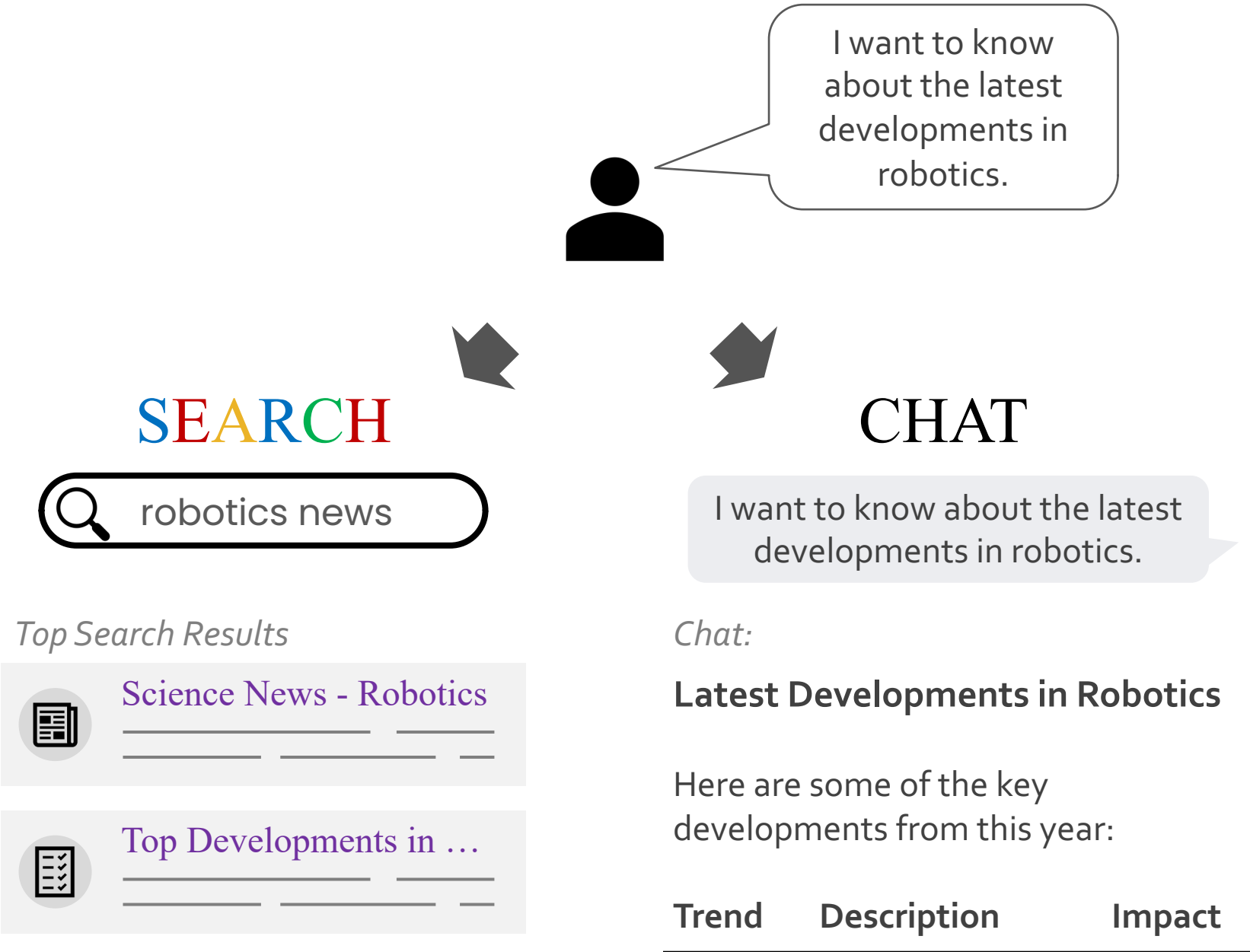


Science News - Robotics



Top Developments in ...

Interpretability for Retrieval



Interpretability for Retrieval

CHAT

I want to know about the latest
developments in robotics.



LLM - Large
Transformer based
Language Model

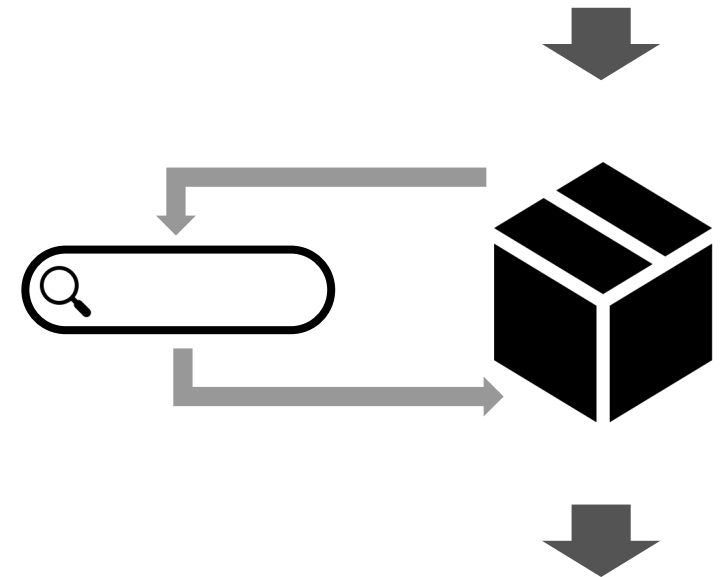


Latest Developments in Robotics
Here are some of the key...

Interpretability for Retrieval

CHAT

I want to know about the latest
developments in robotics.



Latest Developments in Robotics
Here are some of the key...

Interpretability for Retrieval

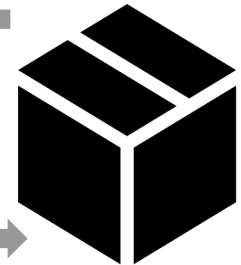
SEARCH



Science News - Robotics

CHAT

I want to know about the latest developments in robotics.



Latest Developments in Robotics
Here are some of the key...

Interpretability for Retrieval

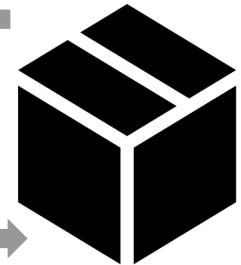
SEARCH



Science News - Robotics

CHAT

I want to know about the latest developments in robotics.



Latest Developments in Robotics
Here are some of the key...

Interpretability for Retrieval

SEARCH



Science News - Robotics

Interpretability for Information Retrieval

- How does the model determine **relevance** of an article?
- How to **improve** the relevance ranking?
- Can we **manipulate** the ranking?

Talk Outline

Interpretability for Generative Retrieval

Model-Editing for Generative Retrieval

Ongoing Projects

Talk Outline

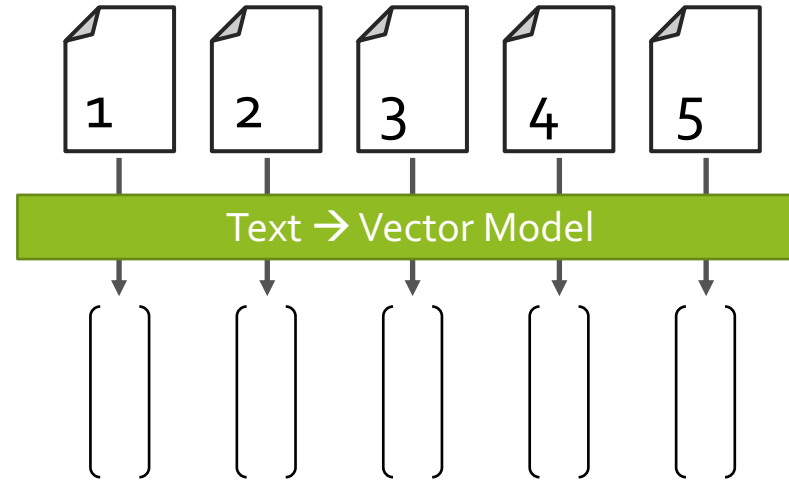
Interpretability for Generative Retrieval

Model-Editing for Generative Retrieval

Ongoing Projects

Current Retrieval Systems

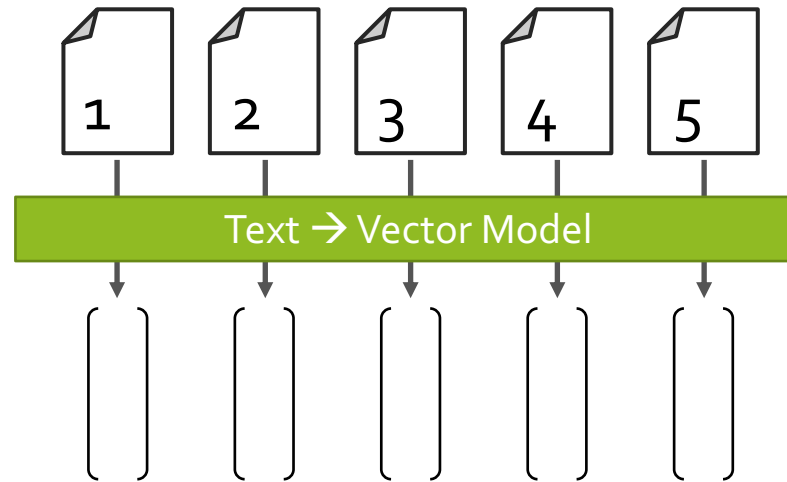
Dense Retrieval



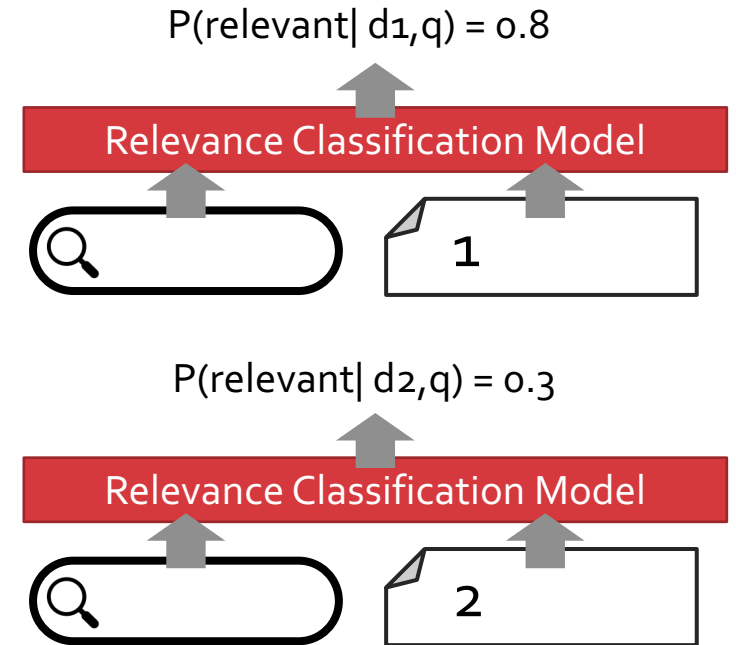
Cross-Encoders

Current Retrieval Systems

Dense Retrieval

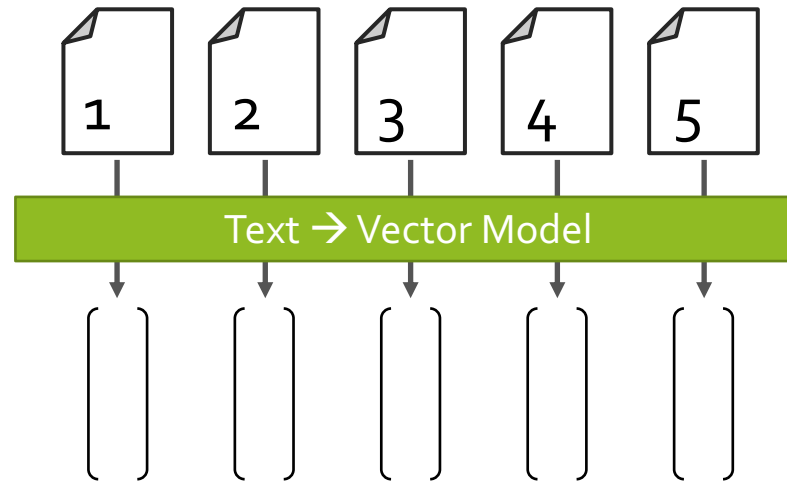


Cross-Encoders

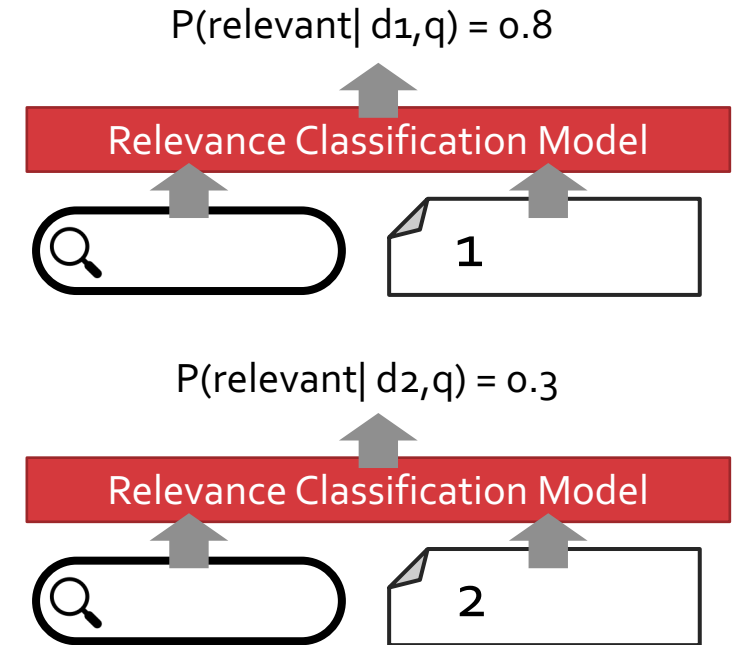


Current Retrieval Systems

Dense Retrieval



Cross-Encoders

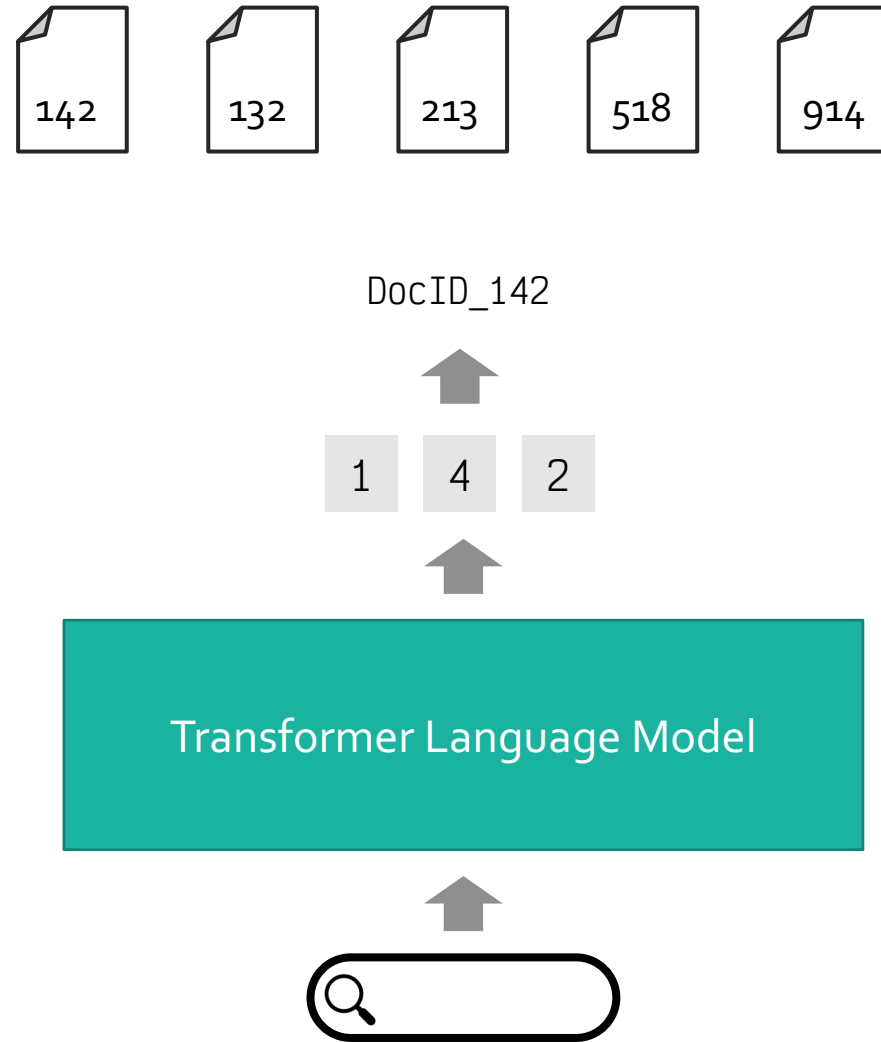


Inference Speed

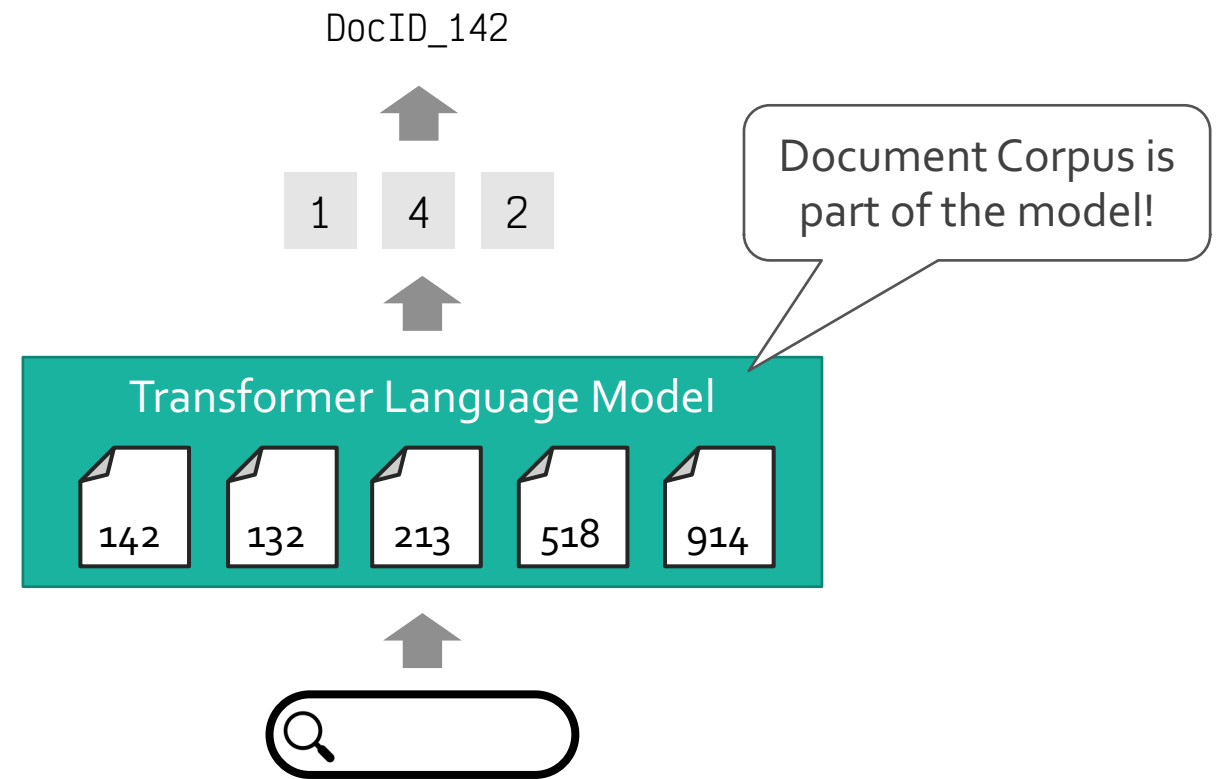
Retrieval Performance

Generative Information Retrieval

Generative Information Retrieval



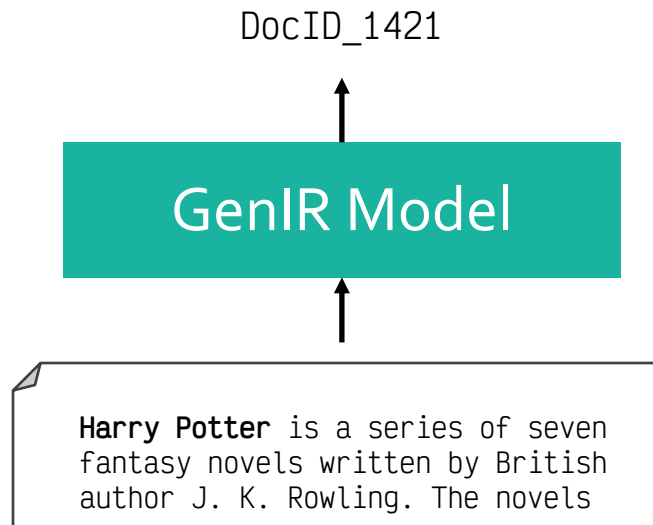
Generative Information Retrieval



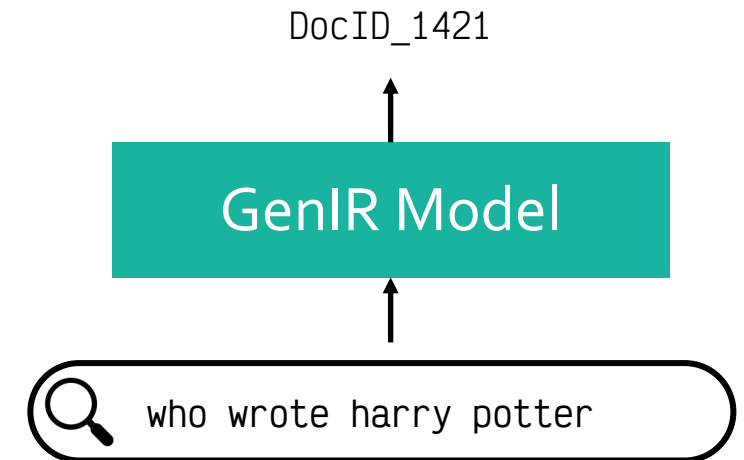
GenIR Training

Goal: Associate the content of a document with its ID

Indexing Phase



Retrieval Phase

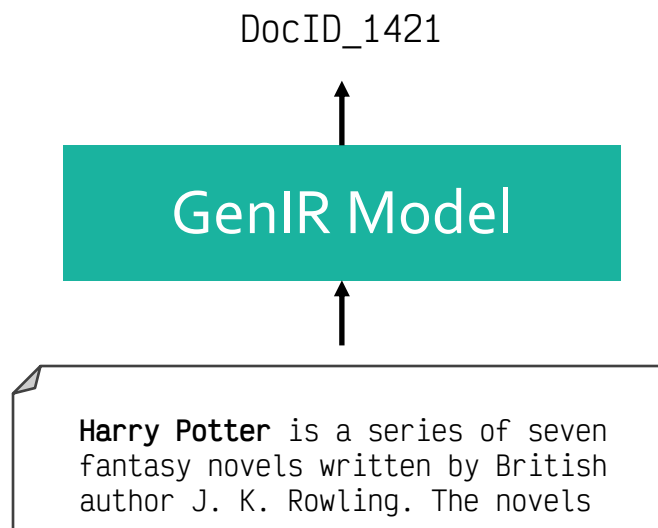


GenIR Training

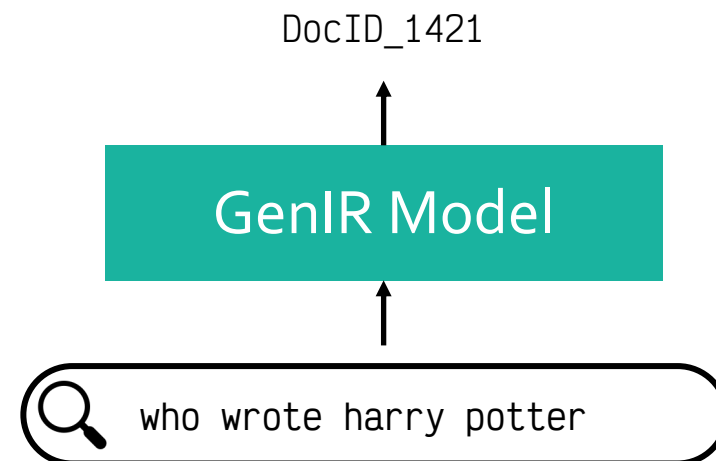
Goal: Associate the content of a document with its ID

Problem: Corpus is “hard-coded” as part of the model.

Indexing Phase

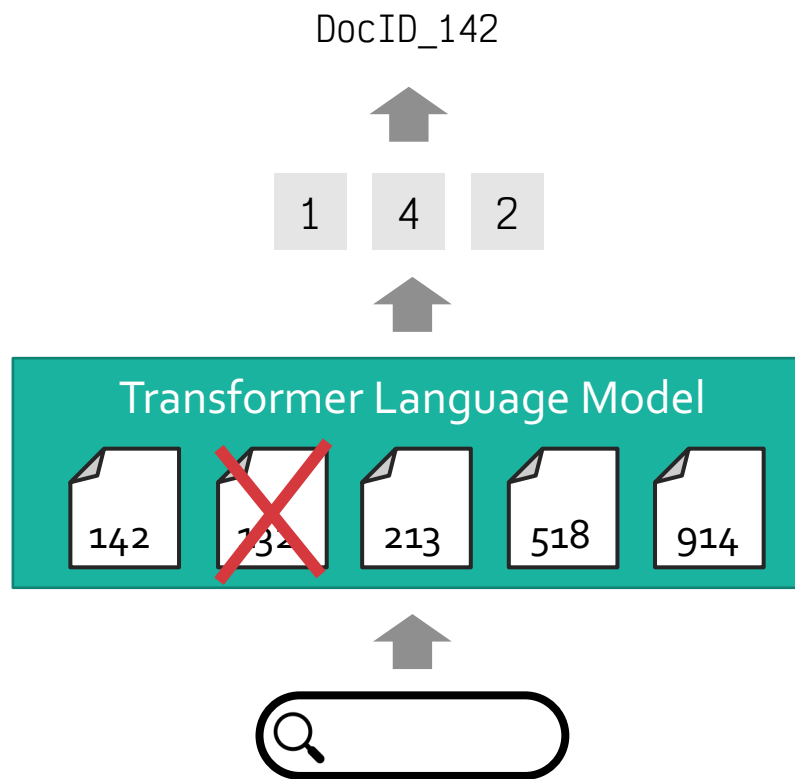


Retrieval Phase



The Problem

Delete Documents

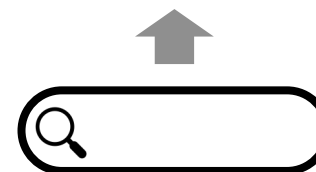
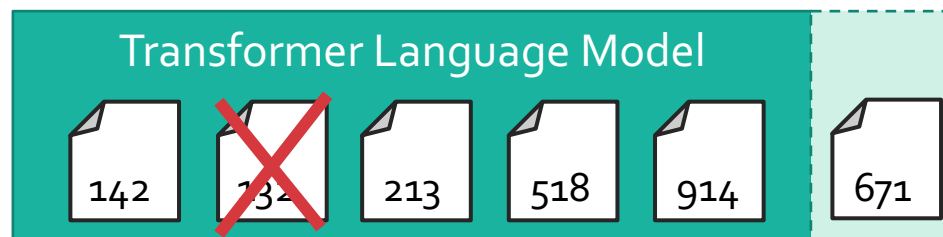
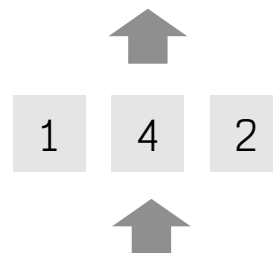


The Problem

Delete Documents

Add Documents

DocID_142



The Problem

Delete Documents

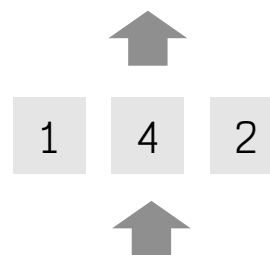
Add Documents

Change Documents

The state-of-the-art
for robotic movement
is based on ~~diffusion~~.

VLAs

DocID_142



Transformer Language Model



My Research

Delete Documents

Add Documents

Change Documents

Goal: Edit the corpus without re-training the model

Challenges

- ☐ Find out how these models work.
- ☐ Develop editing method.

My Research

Delete Documents

Add Documents

Change Documents

Goal: Edit the corpus without re-training the model

Challenges

- ☐ Find out how these models work.
- ☐ Develop editing method.

Reverse-Engineering the Retrieval Process in GenIR Models

Anja Reusch & Yonatan Belinkov – SIGIR 2025



Funding:



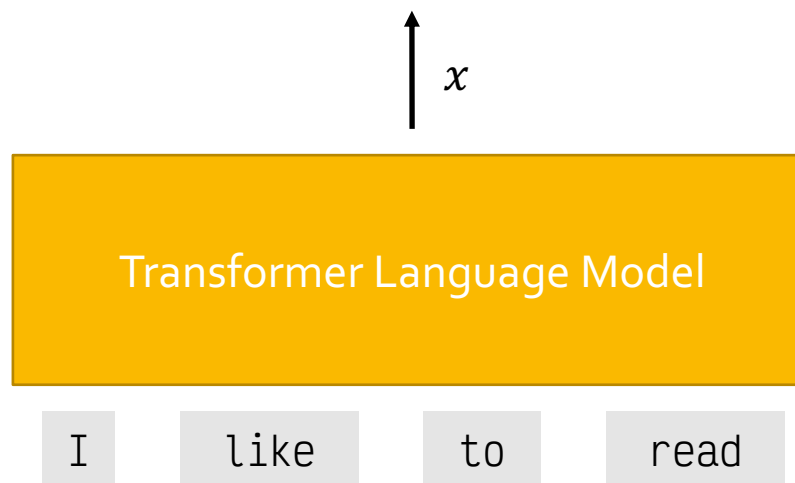
Funded by
the European Union



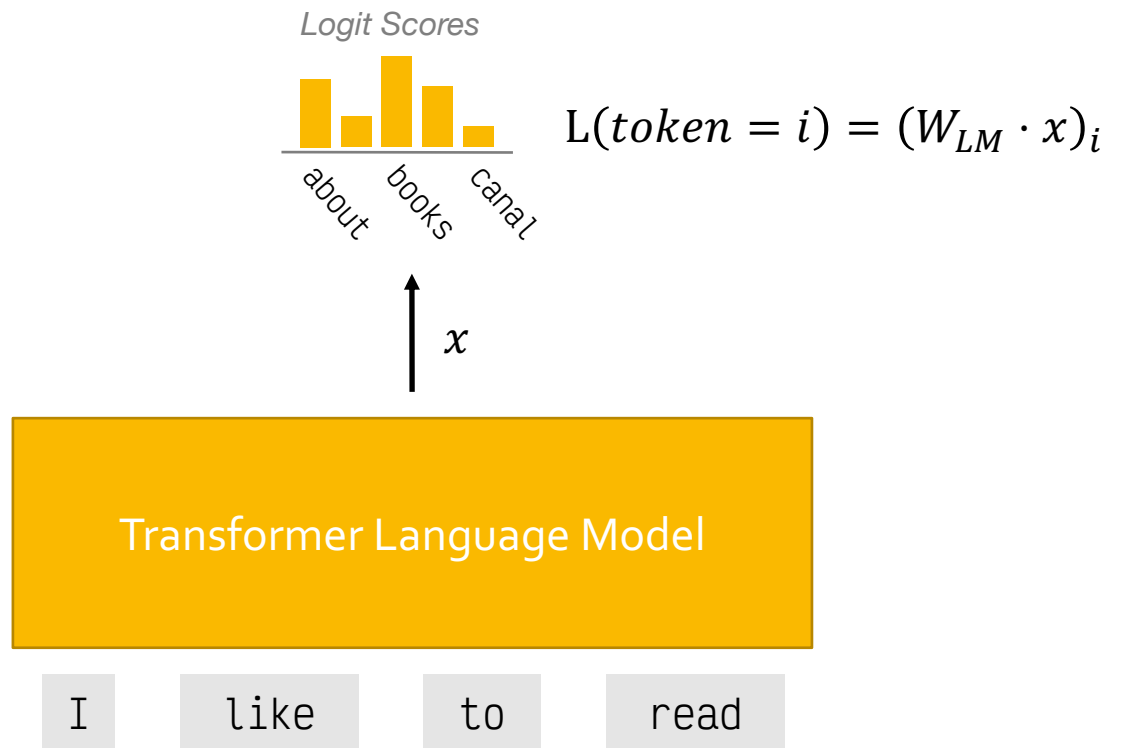
European Research Council
Established by the European Commission



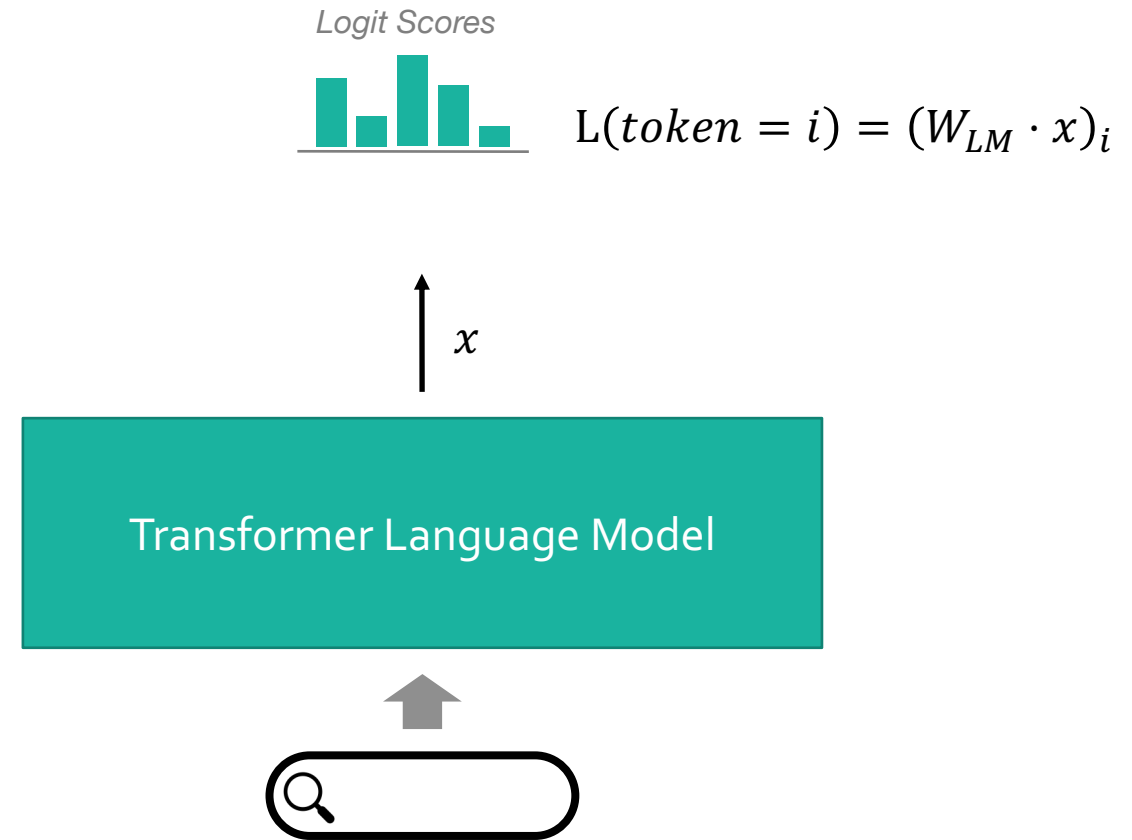
Generative Models



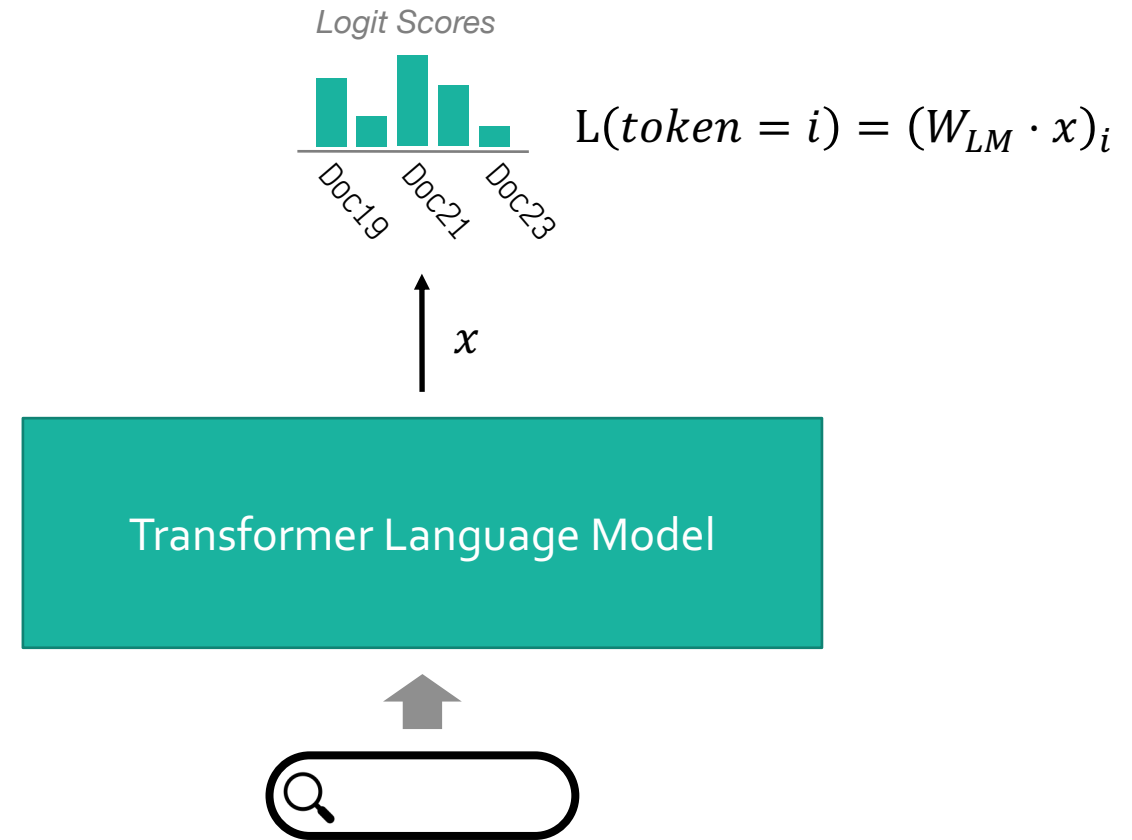
Generative Models



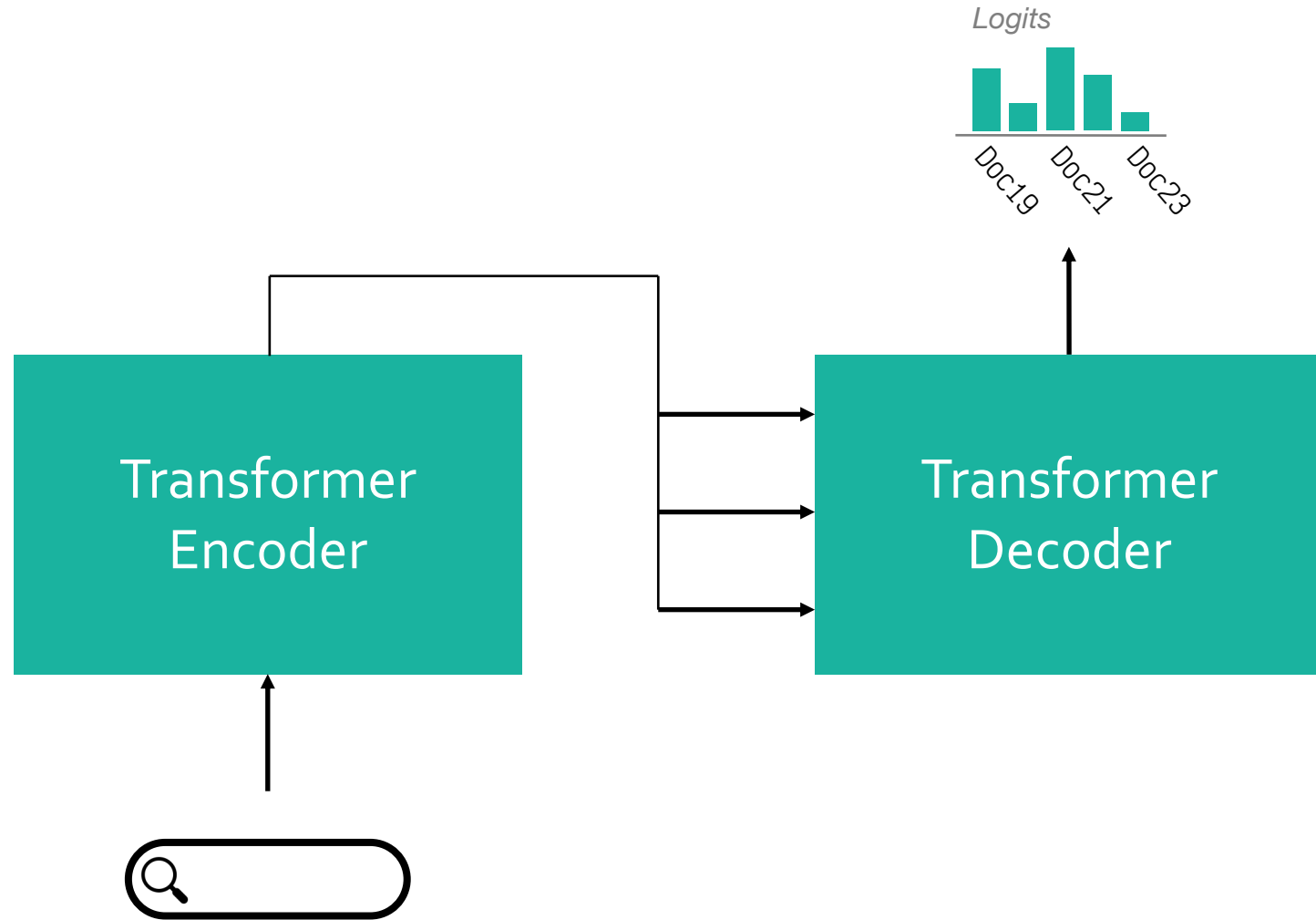
Generative Information Retrieval



Generative Information Retrieval

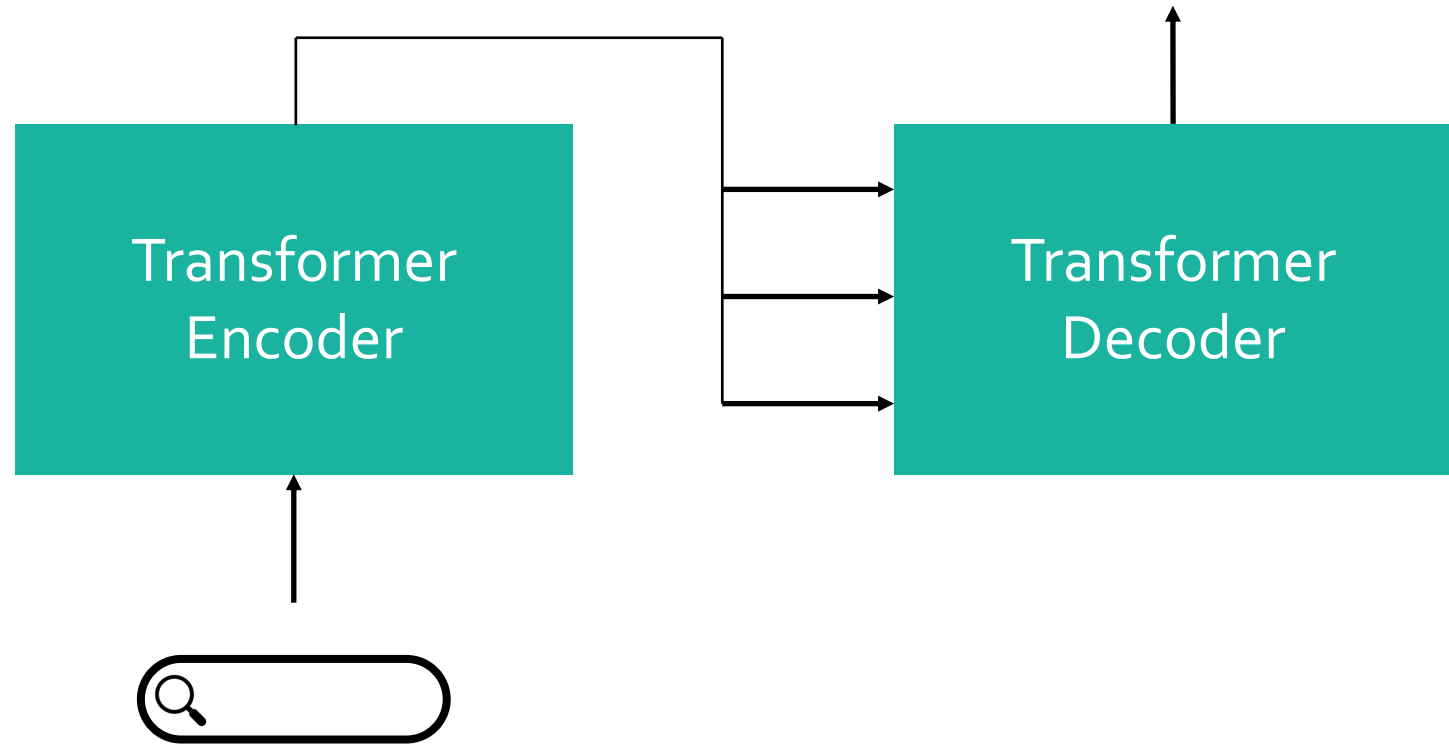


Generative Encoder- Decoder

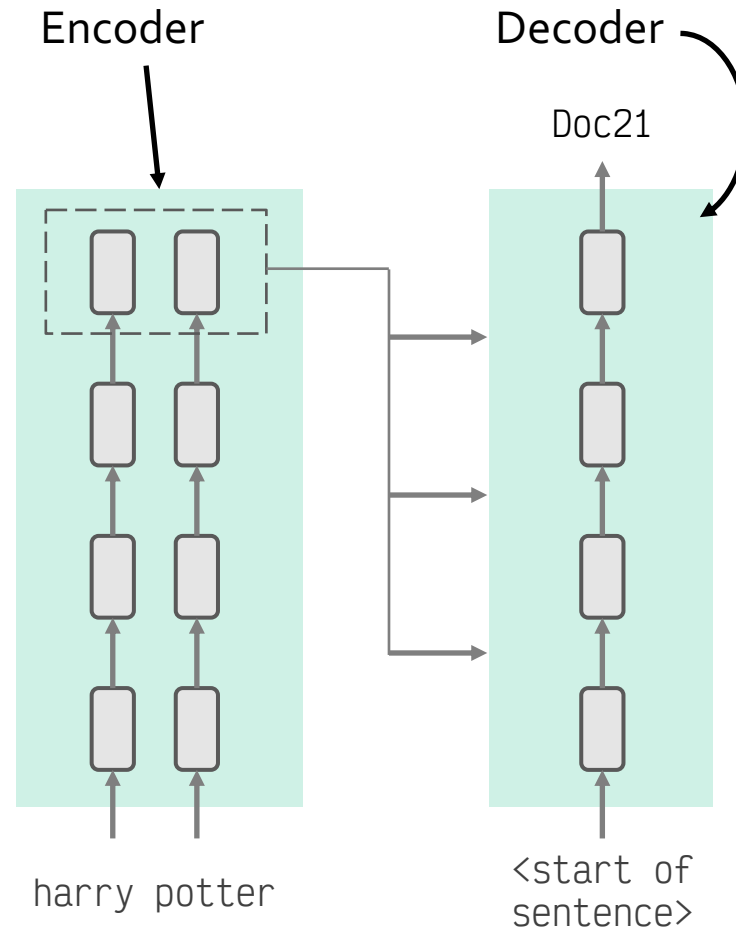


Generative Encoder- Decoder

Pre-Trained T5 +
Fine-Tuning as GenIR Model

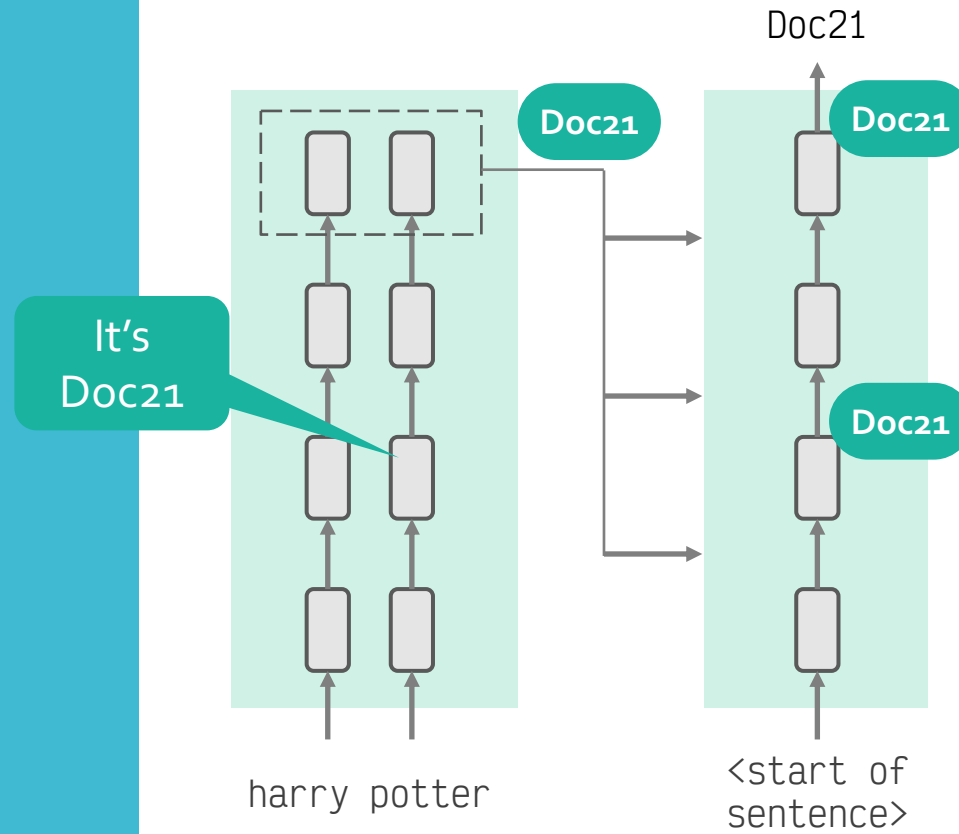


Which part identifies the document?



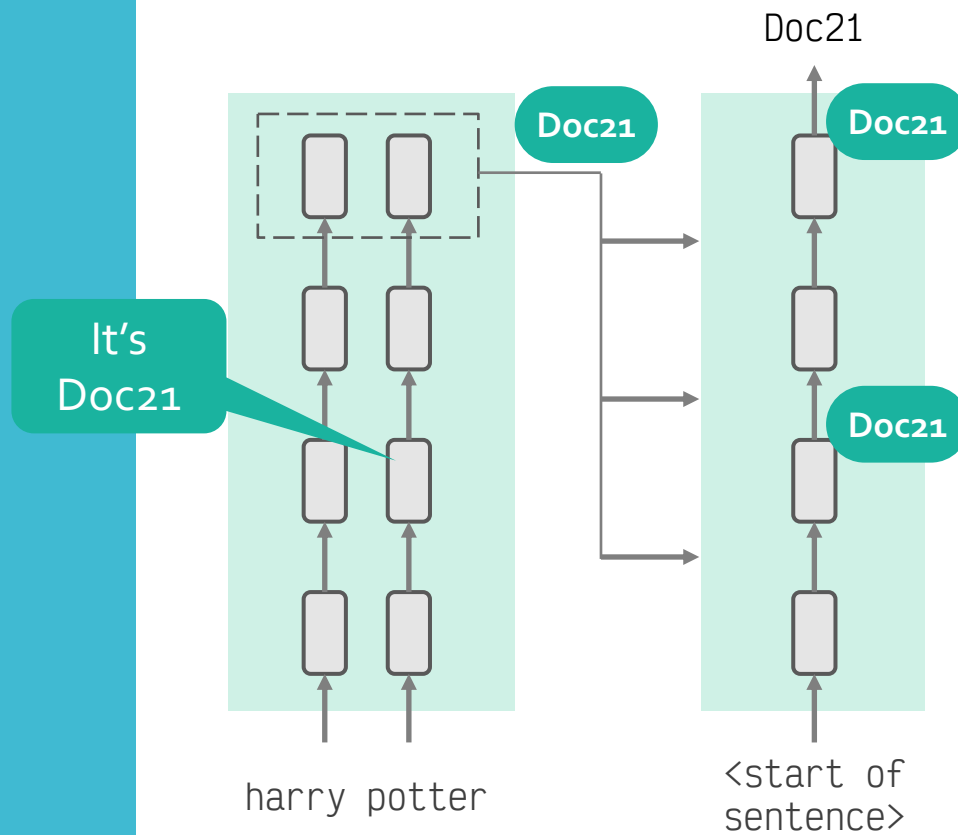
Which part identifies the document?

Encoder identifies the document

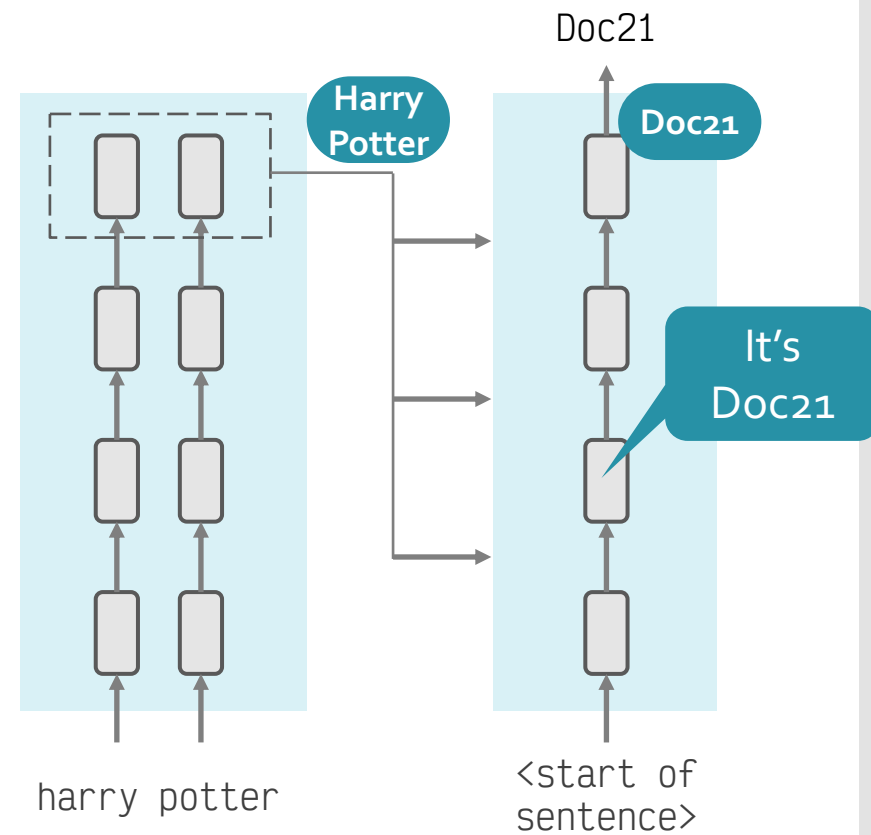


Which part identifies the document?

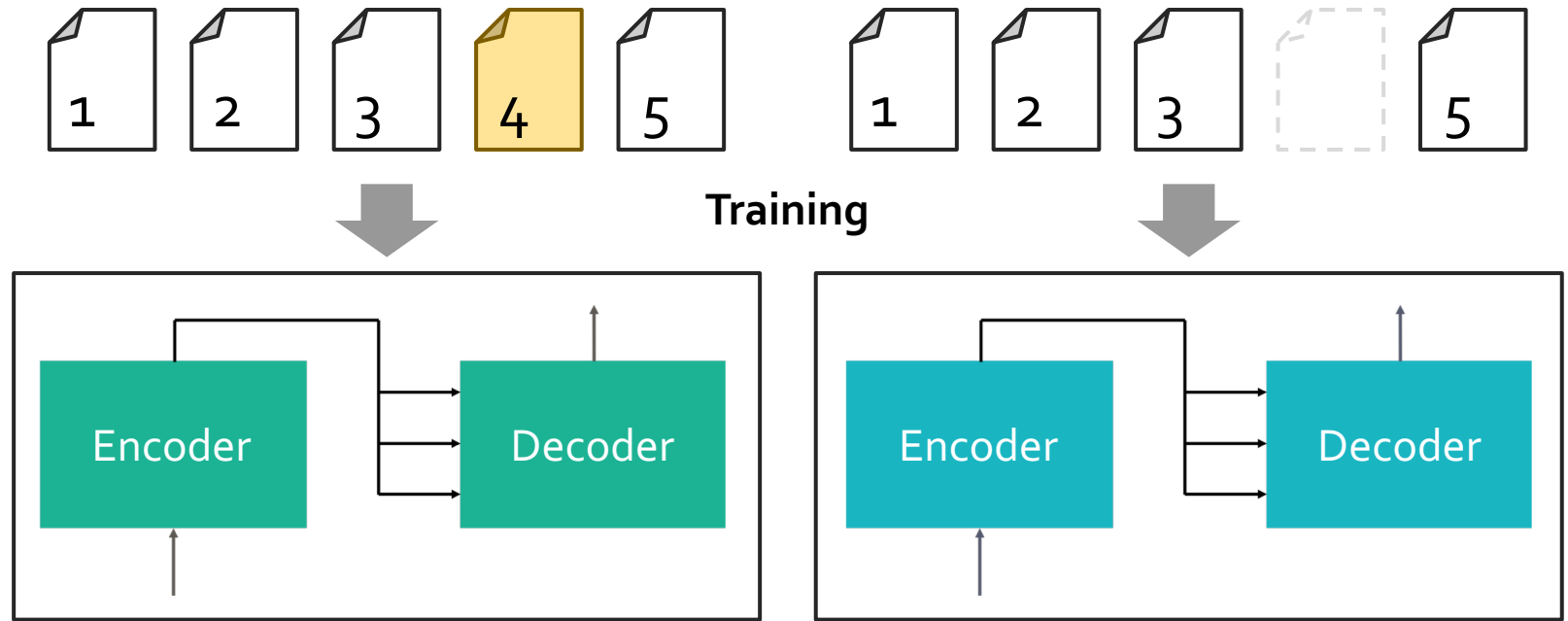
Encoder identifies the document



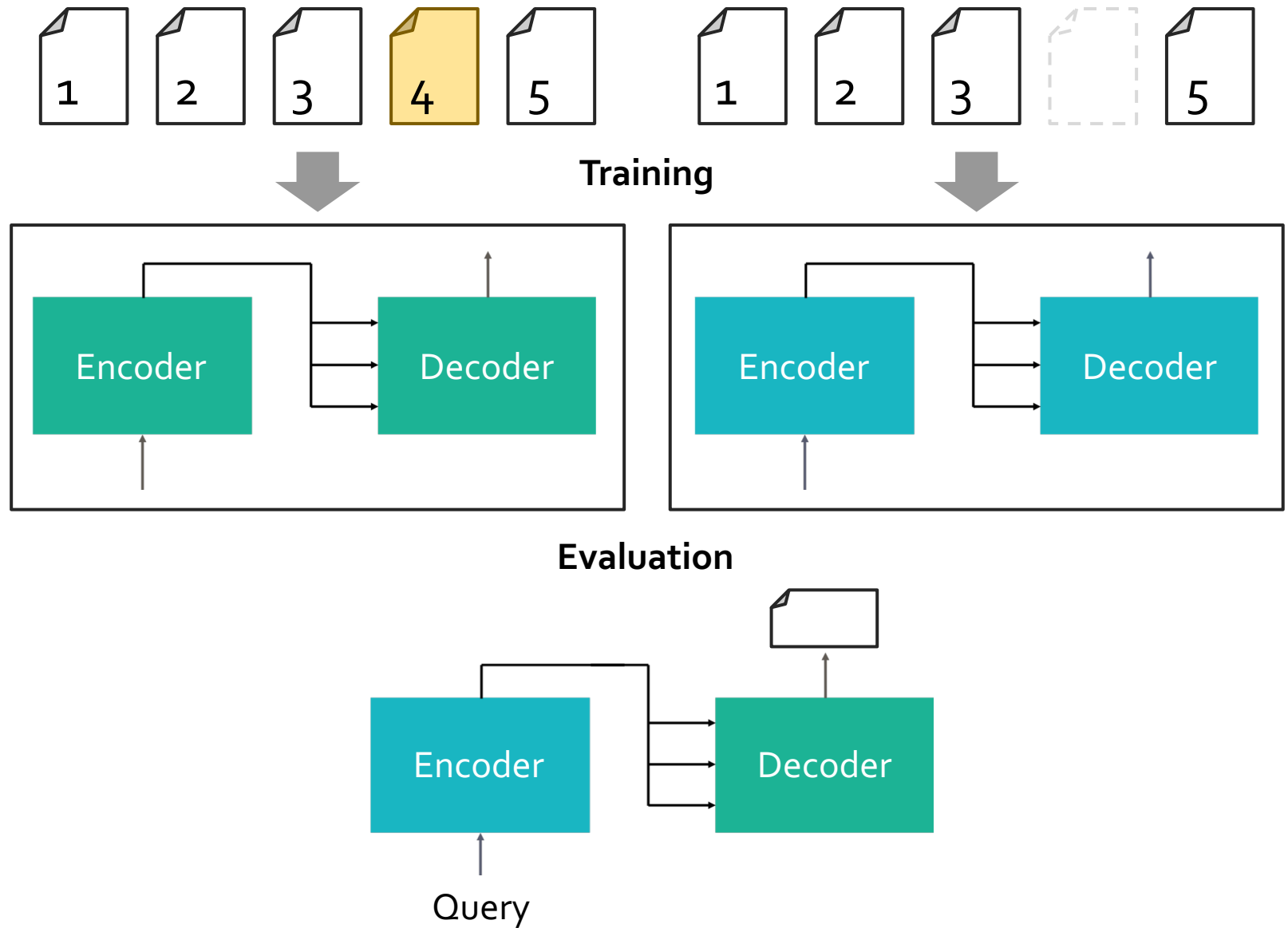
Decoder identifies the document



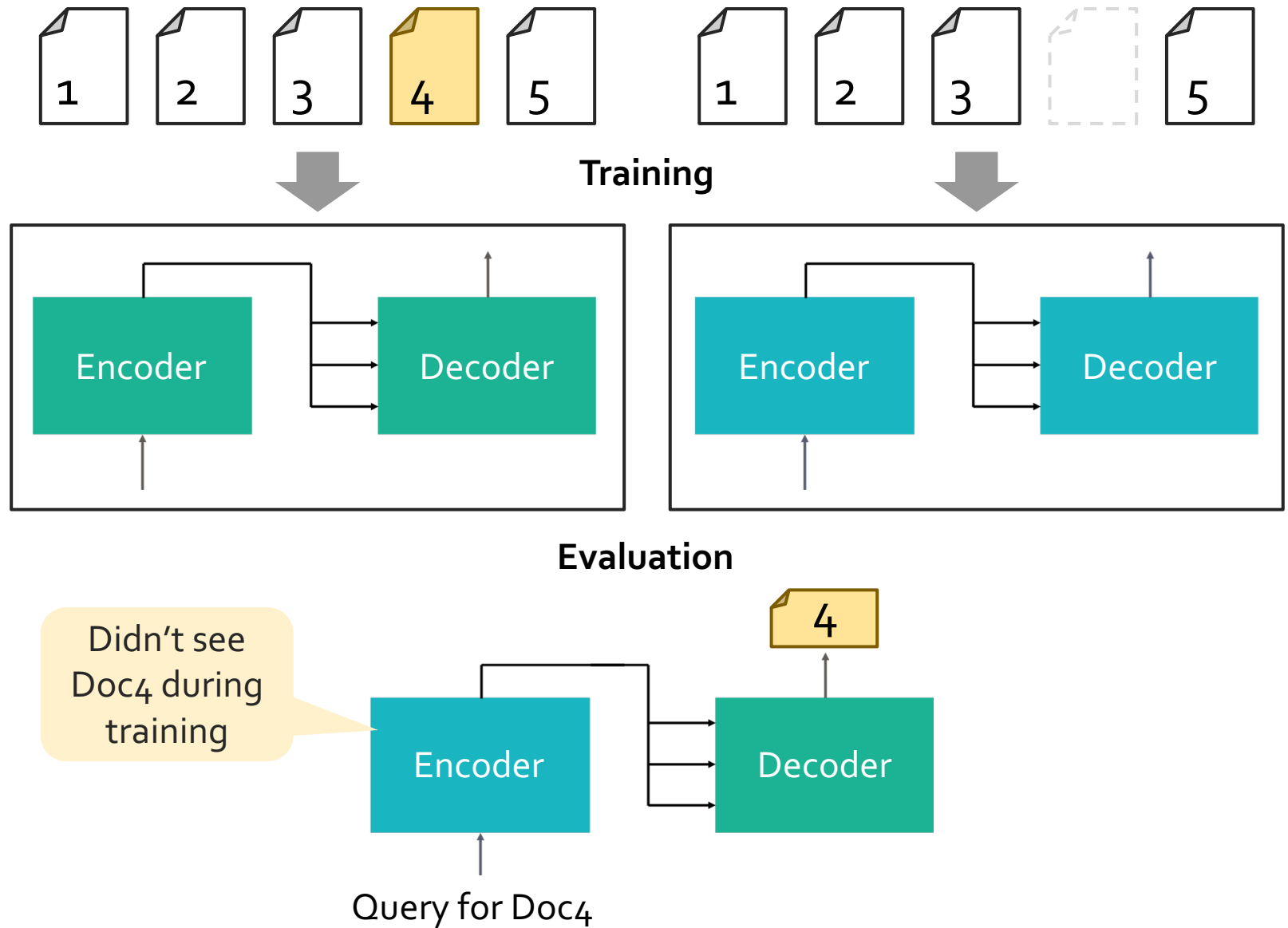
Which part identifies the document?



Which part identifies the document?

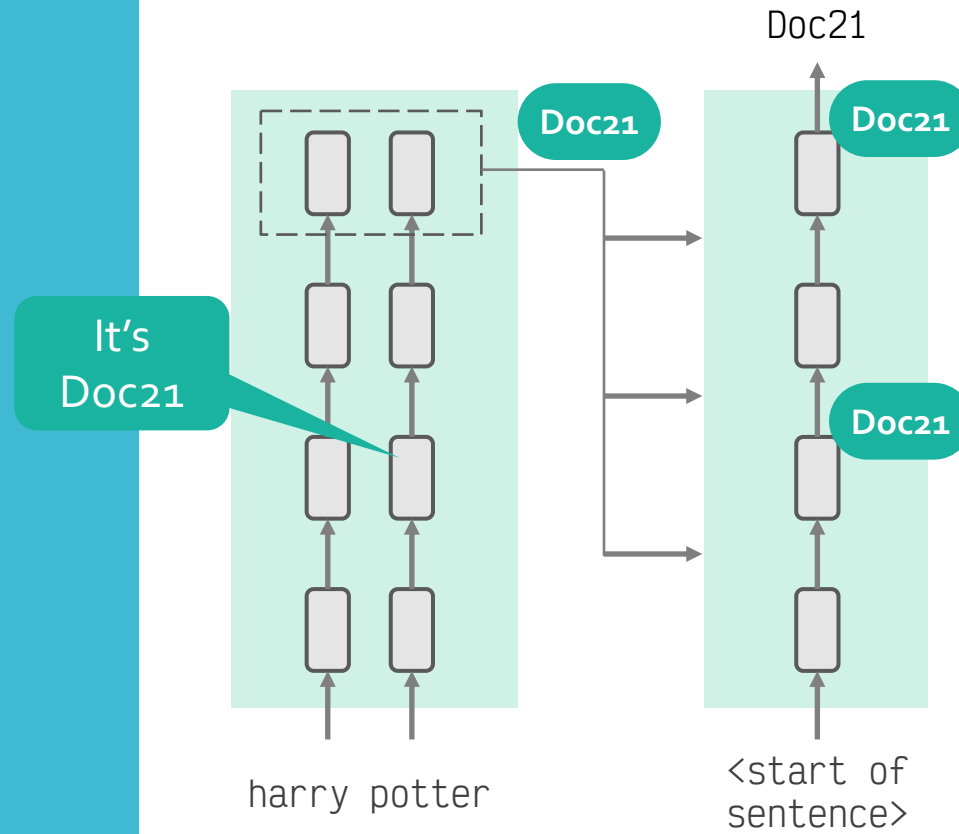


Which part identifies the document?



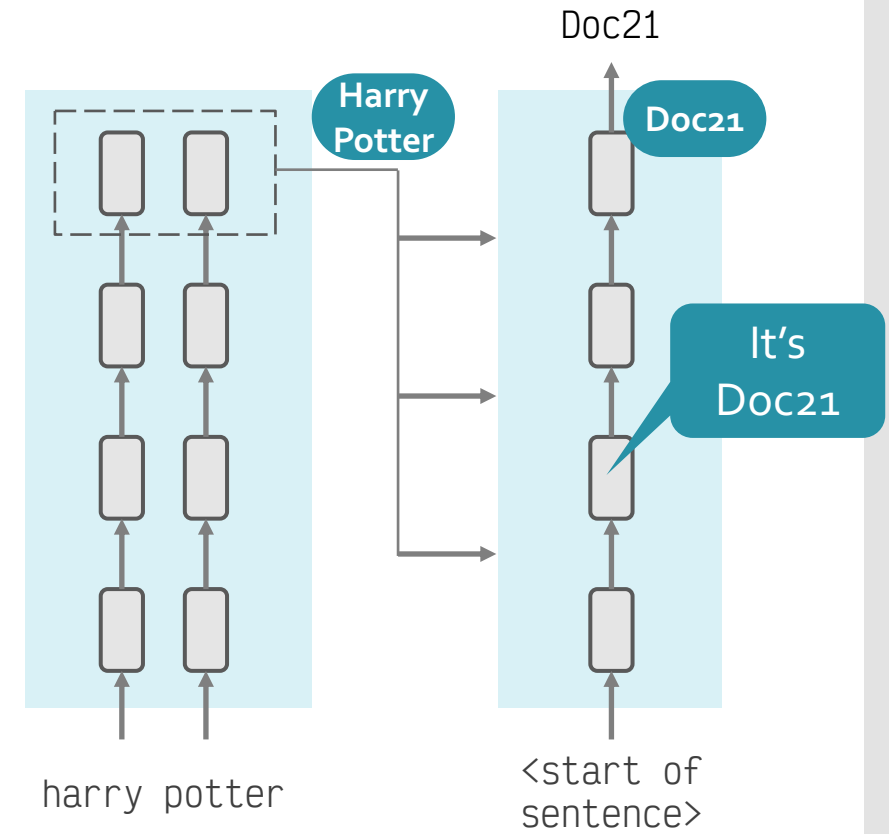
Which part identifies the document?

Encoder identifies the document



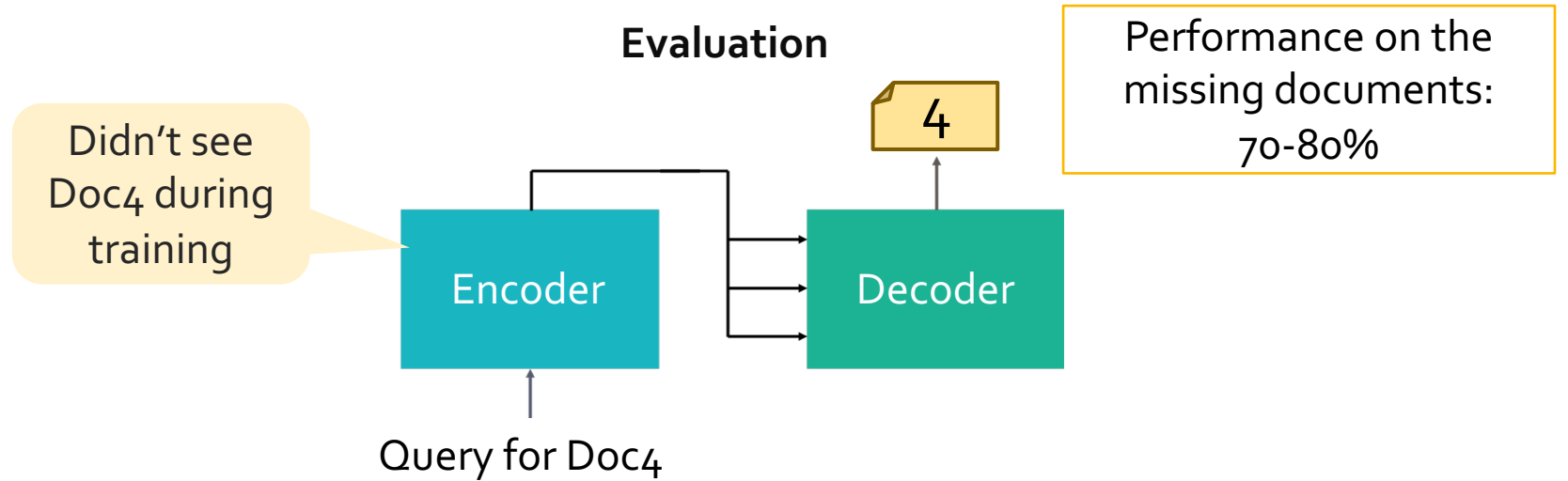
Expectation: 0% Accuracy

Decoder identifies the document



Expectation: ~High Accuracy

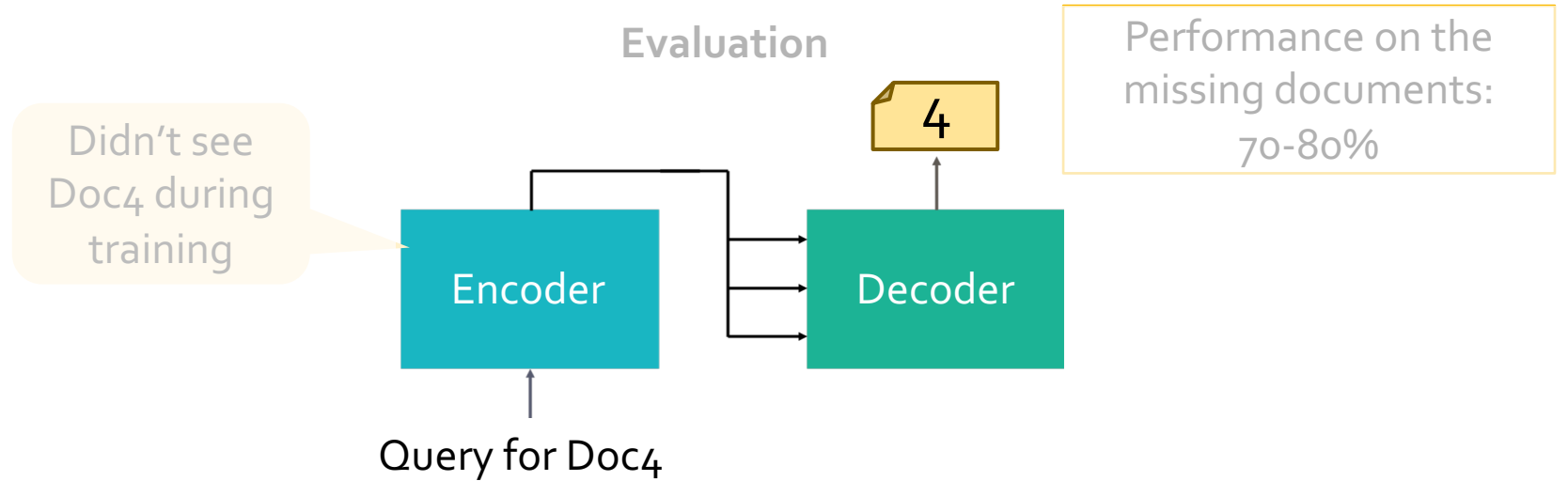
Looking at the Encoder



Setup

- Base Model: Trained on 10k documents
- 4x "Missing Docs Model": Trained on same 9990 documents
- Hits@1 (Correct doc at Rank 1): 70-80%

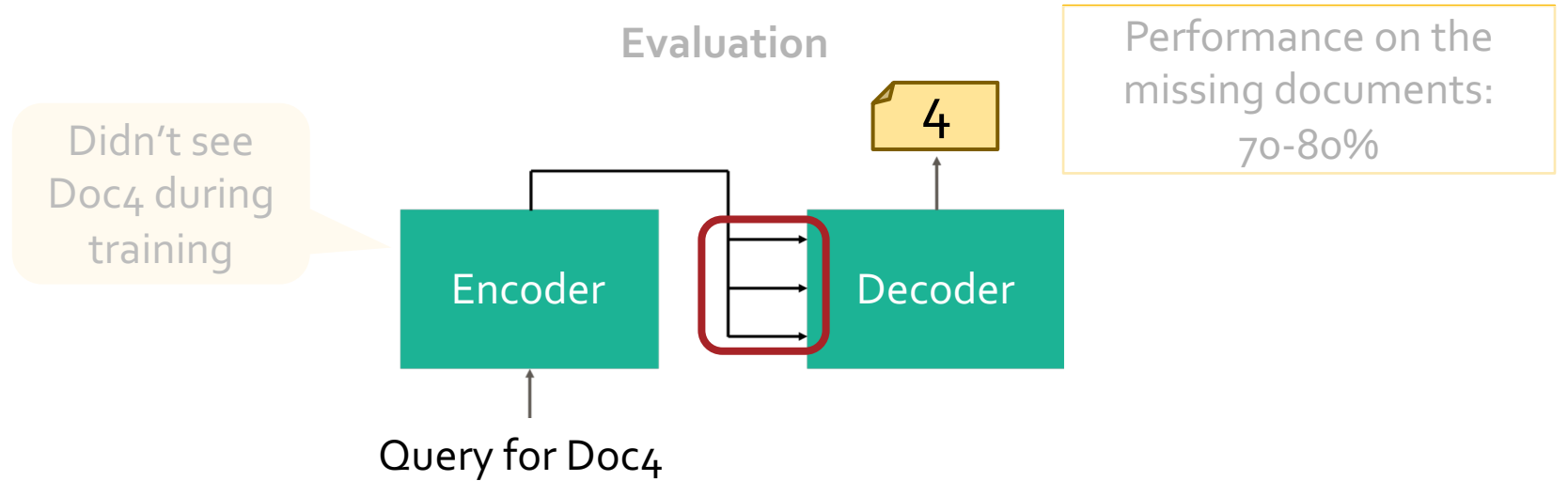
Looking at the Encoder



Conclusion: The decoder identifies the document.

- Base Model: Trained on 10k documents
- 4x "Missing Docs Model": Trained on same 9990 documents
- Correct doc at Position 1: 70-80%

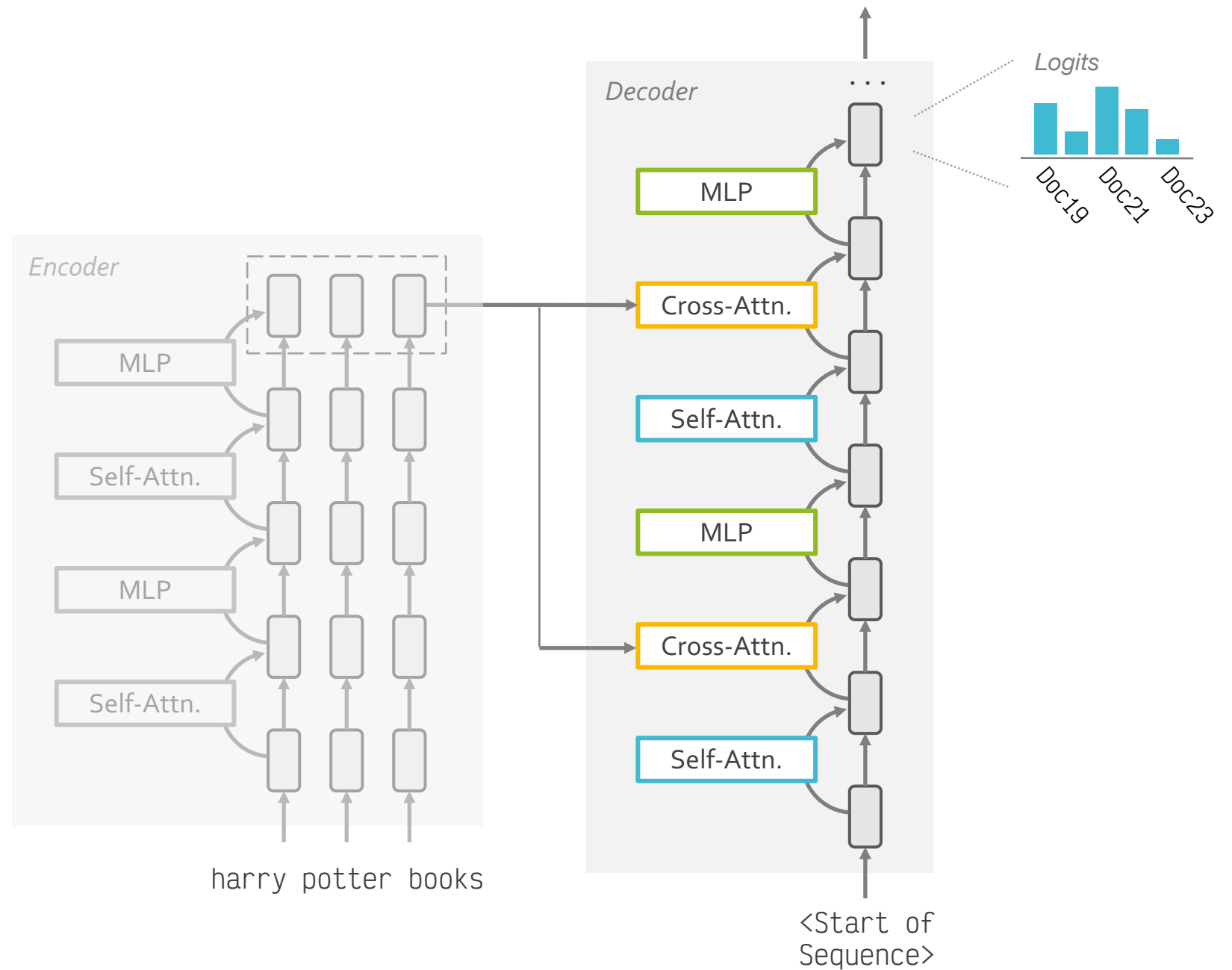
Looking at the Encoder



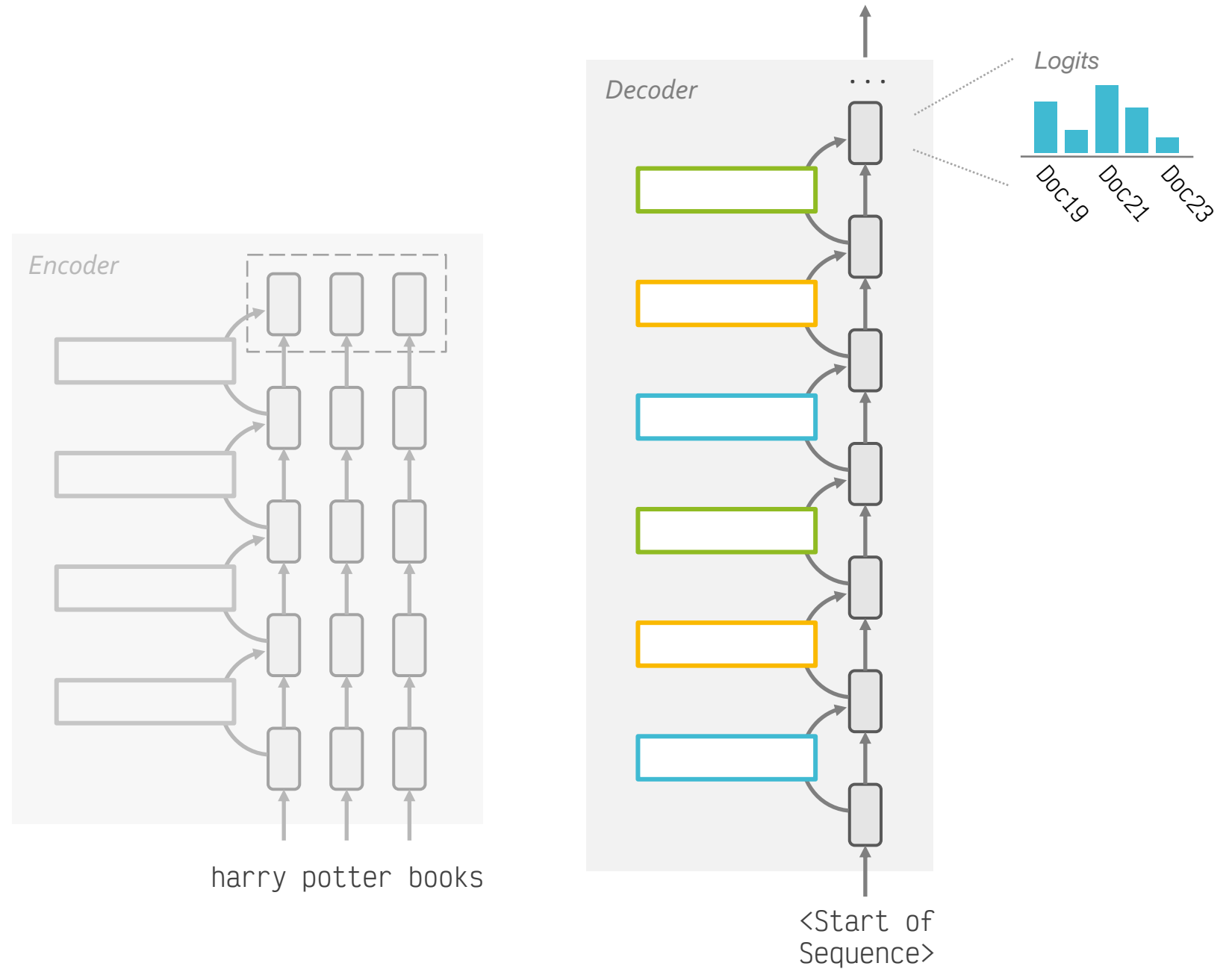
Setup

- Base Model: Trained on 10k documents
- 4x "Missing Docs Model": Trained on same 9990 documents
- Hits@1 (Correct doc at Position 1): 70-80%

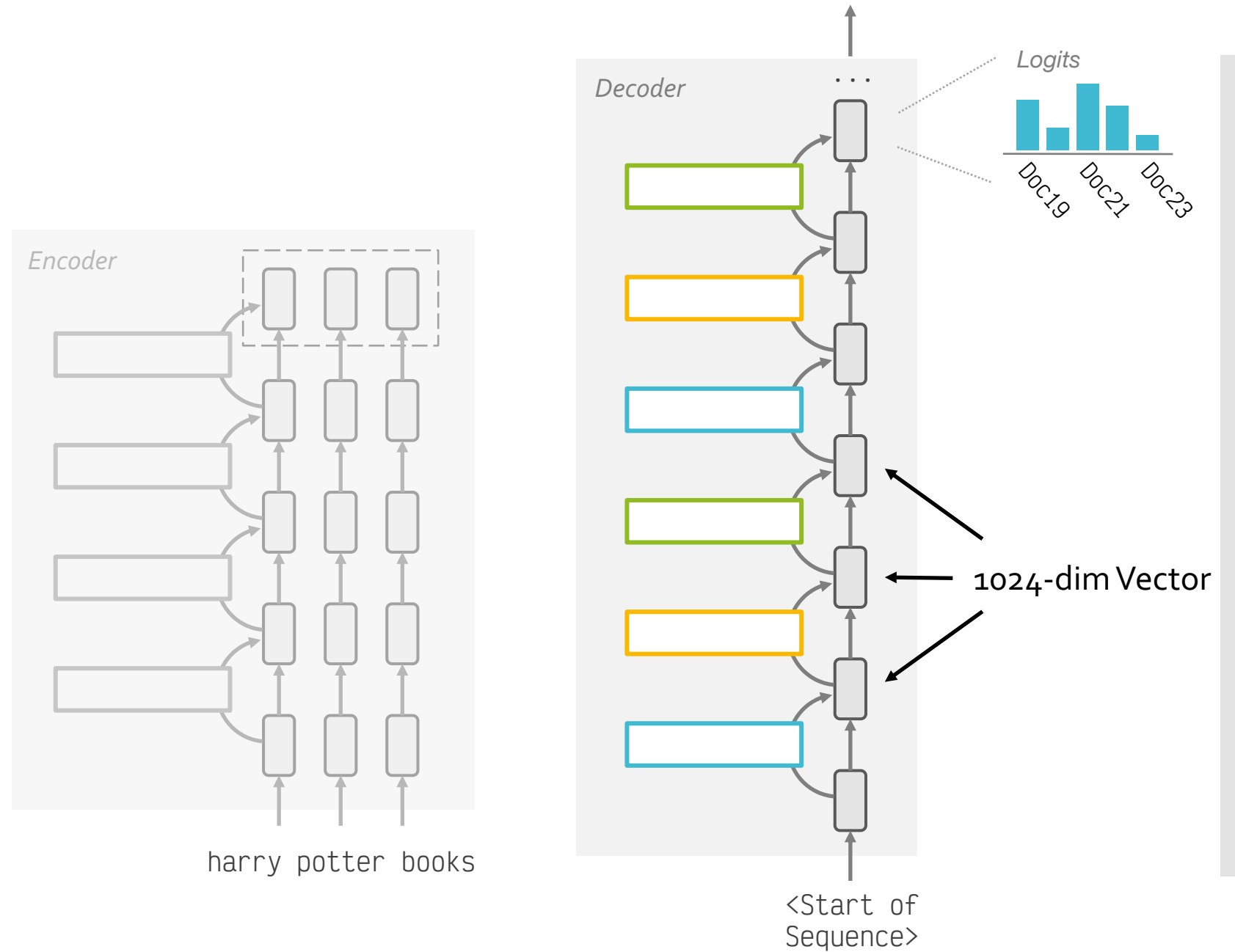
How does a GenIR Model look like?



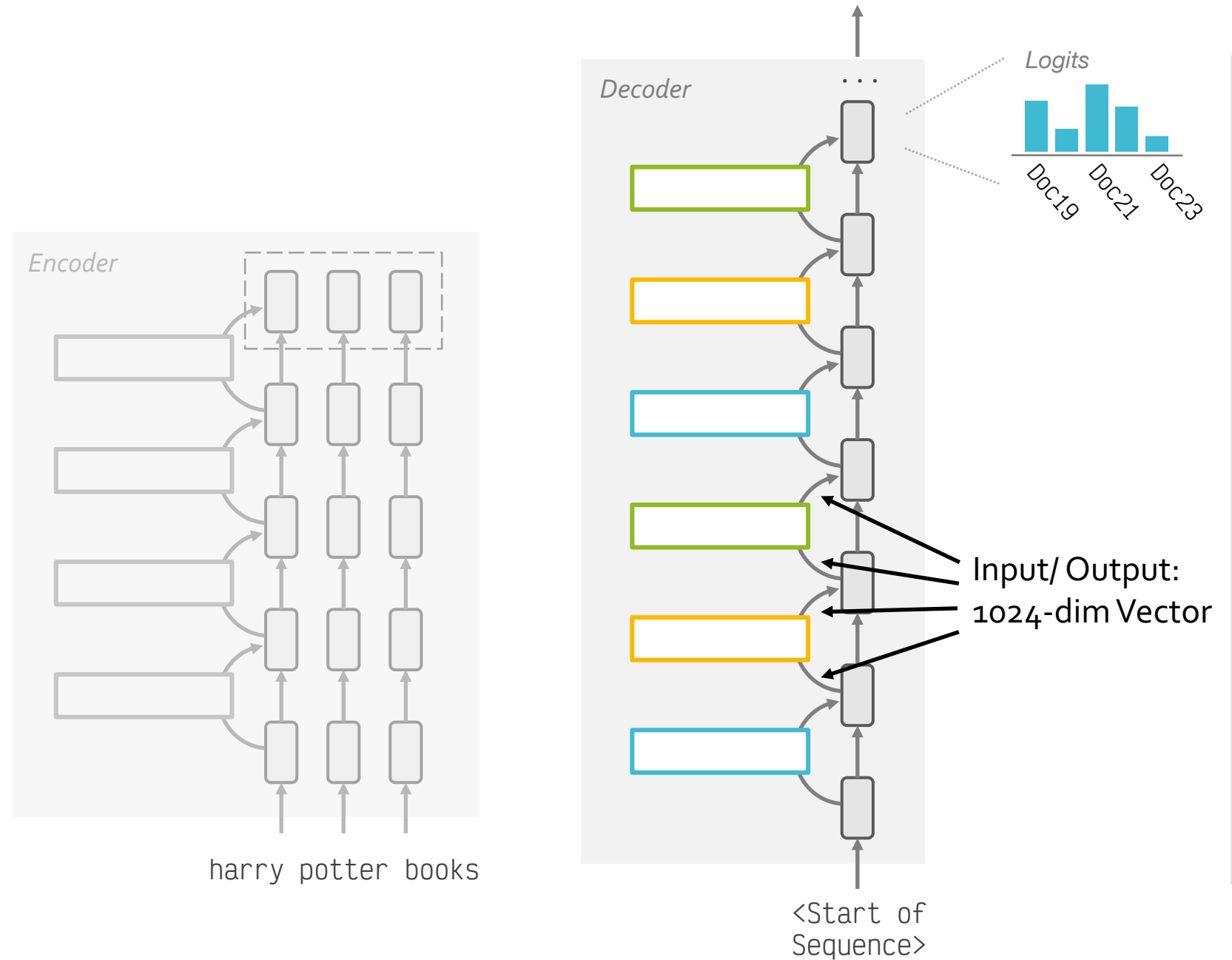
How does a GenIR Model look like?



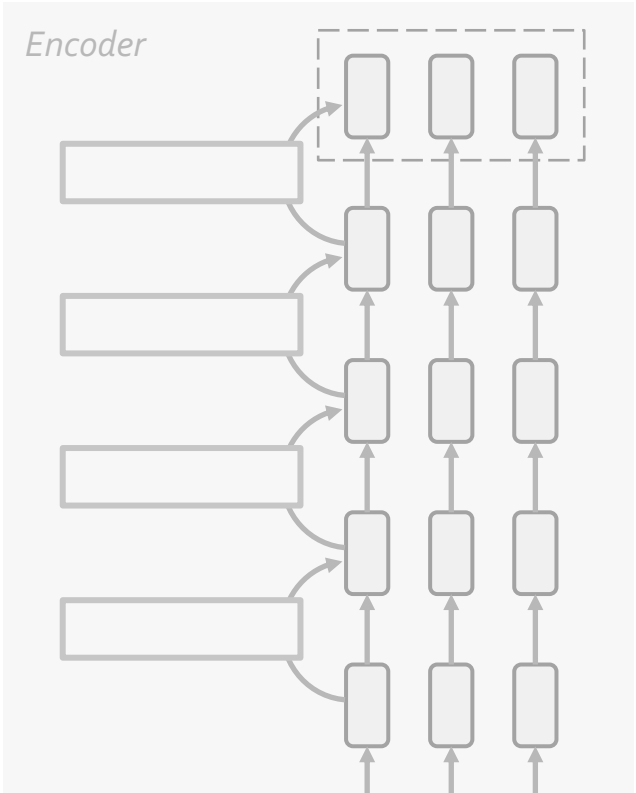
How does a GenIR Model look like?



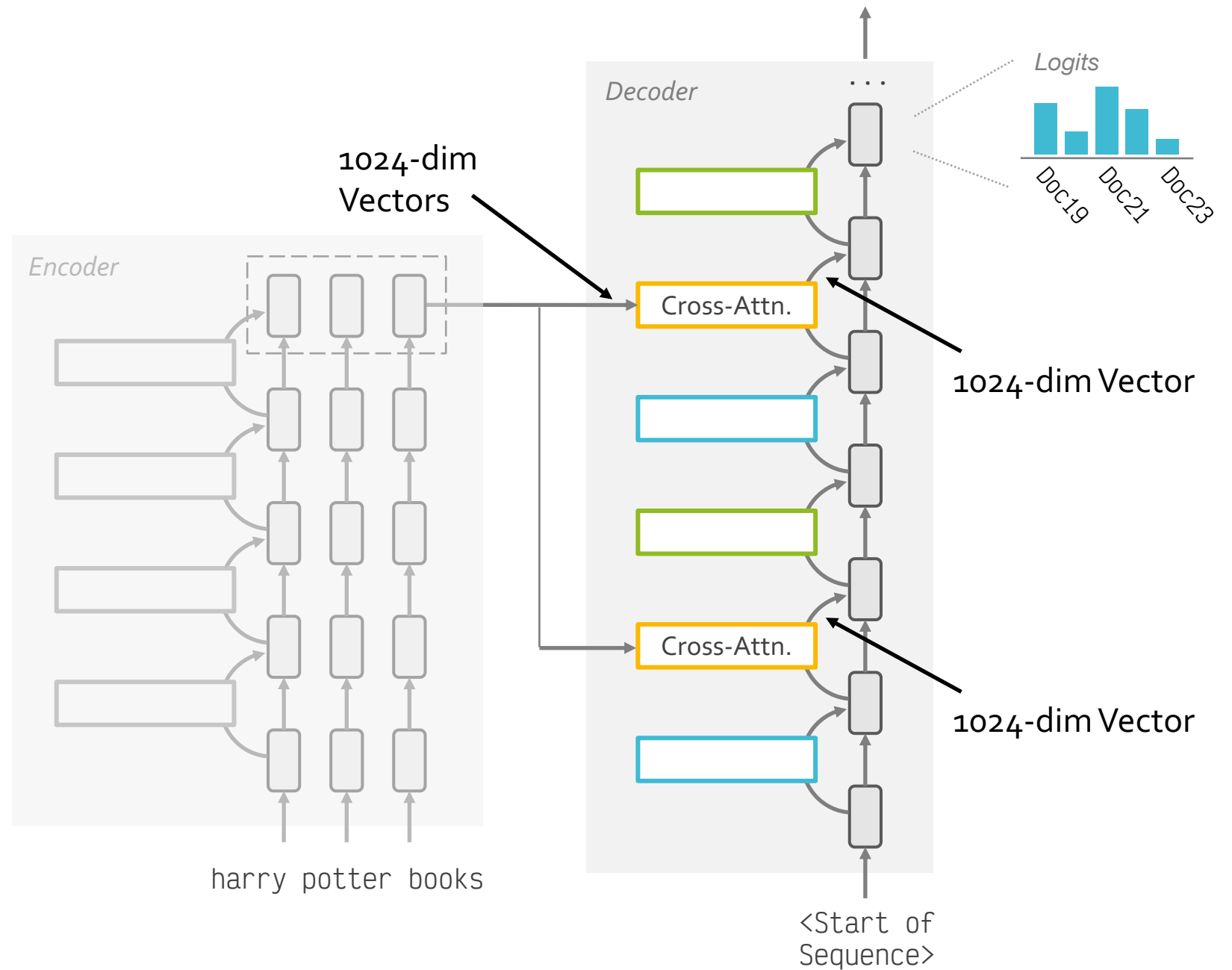
How does a GenIR Model look like?



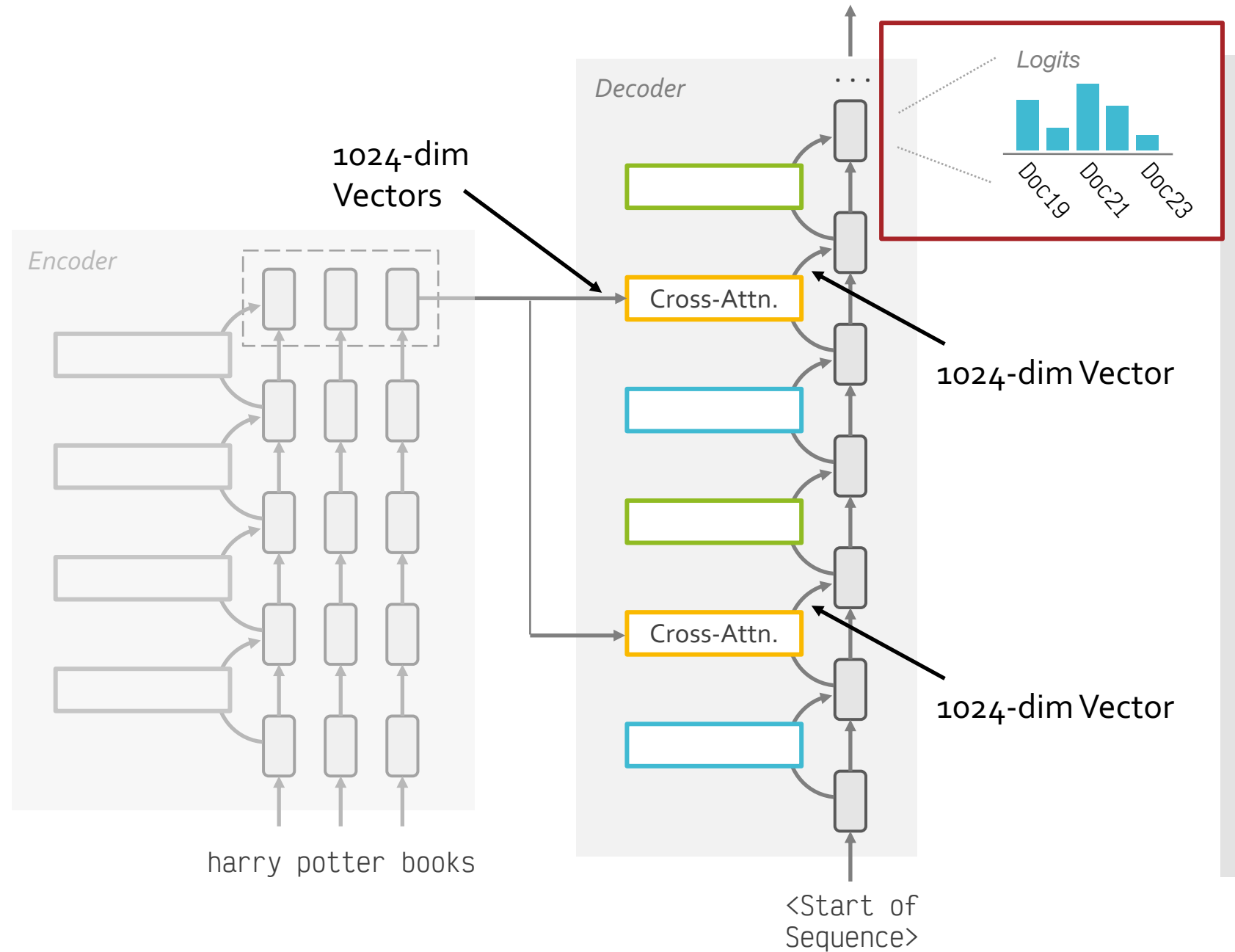
How does a GenIR Model look like?



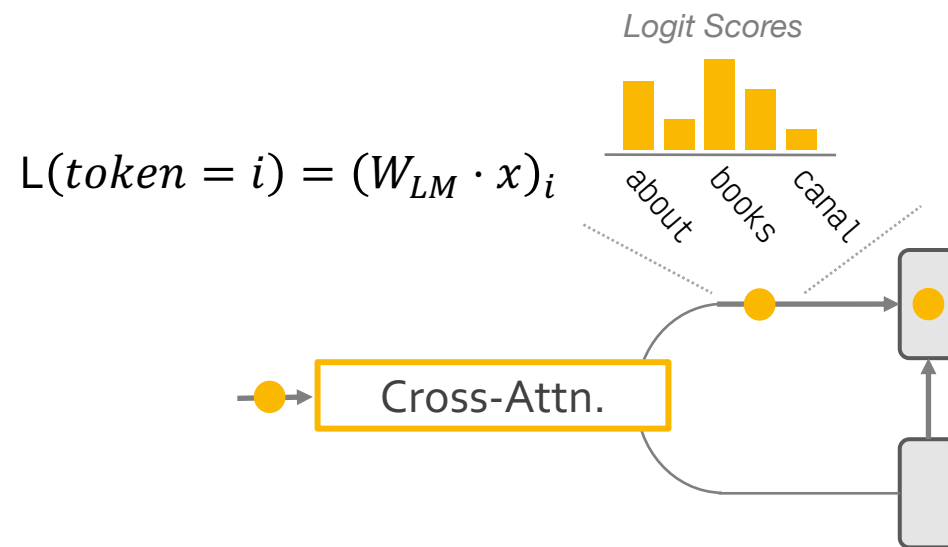
How does a GenIR Model look like?



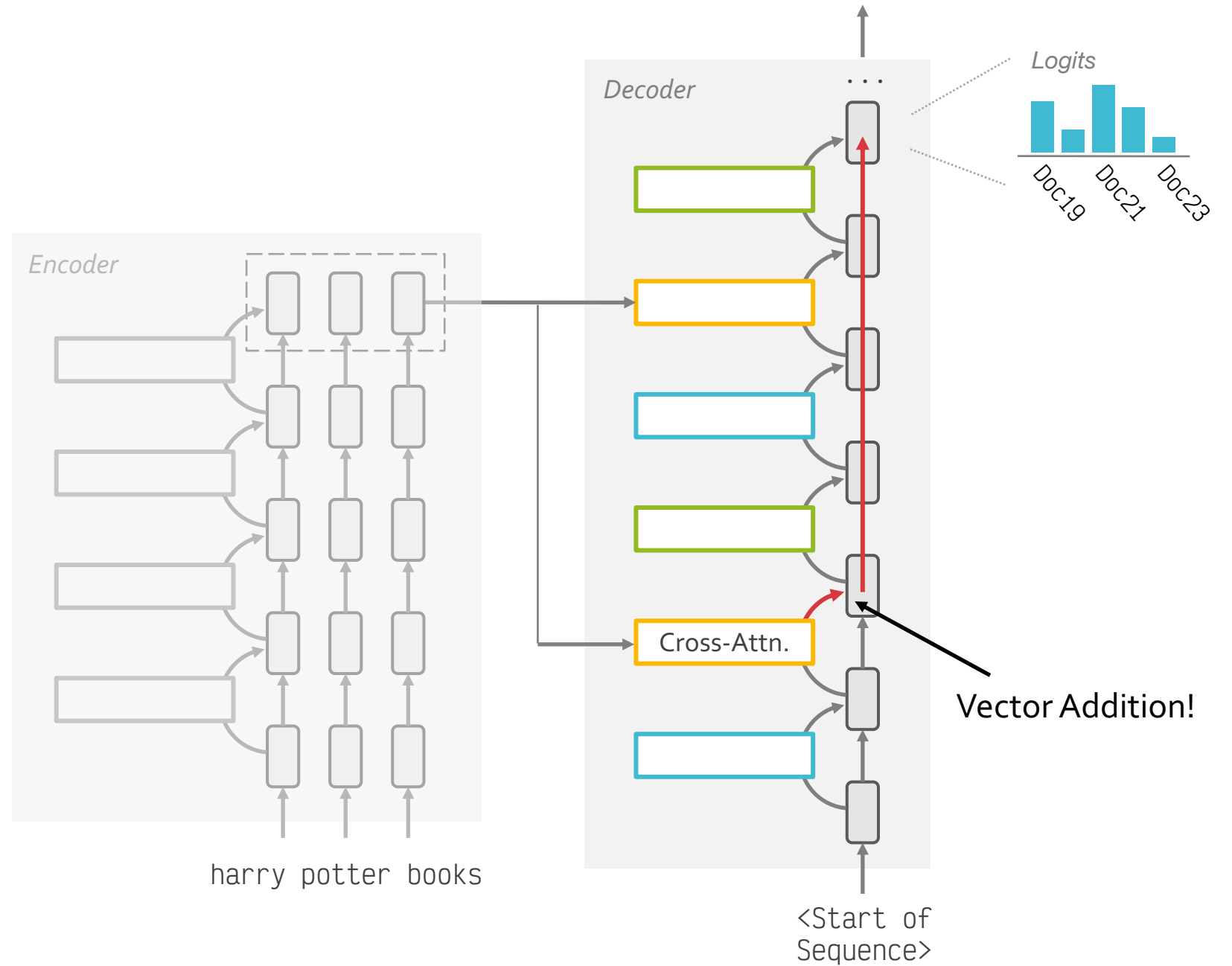
How does a GenIR Model look like?



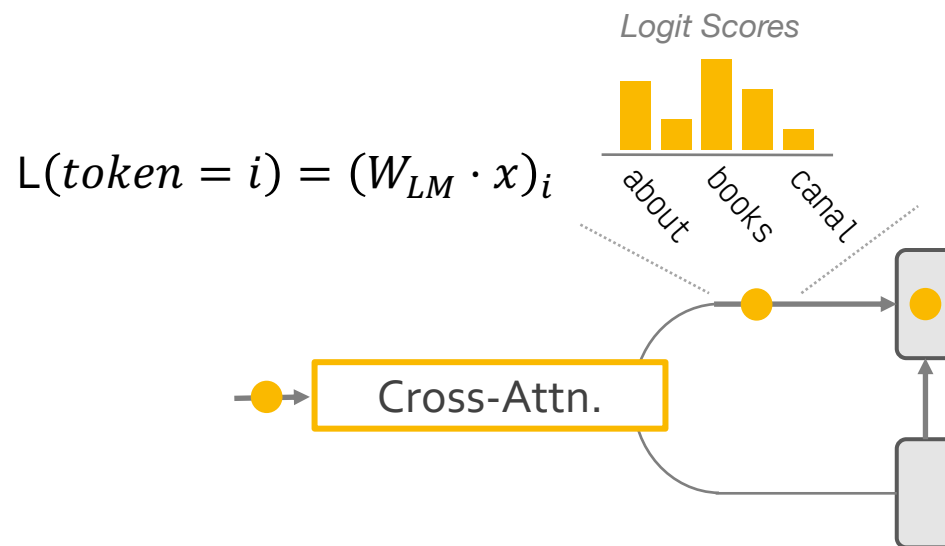
What does
Cross-
Attention
write?



Why does LogitLens work?

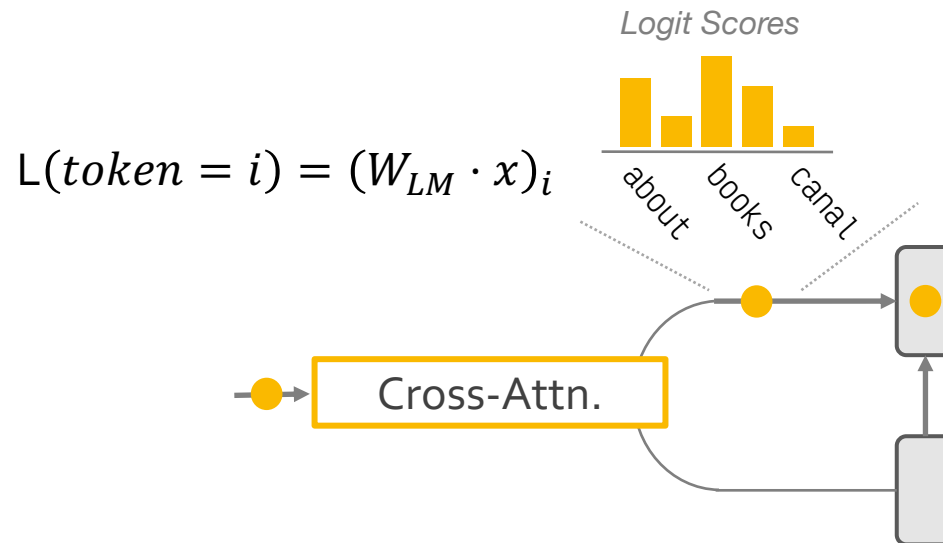


What does Cross- Attention write?



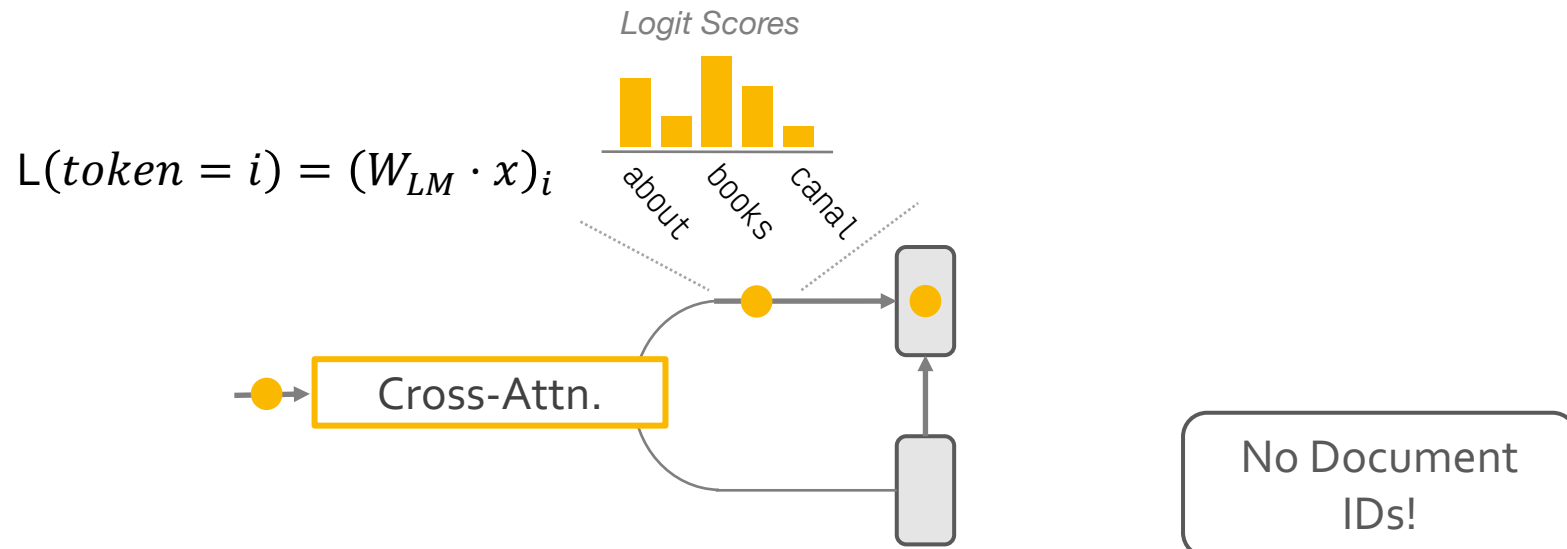
QUERY	LAYER	HEAD	TOP 5 TOKENS
who wrote the harry potter books	Layer 14	Head 2	
	Layer 16	Head 1	
who won the football championship in 2002	Layer 16	Head 1	
	Layer 16	Head 13	

What does Cross- Attention write?



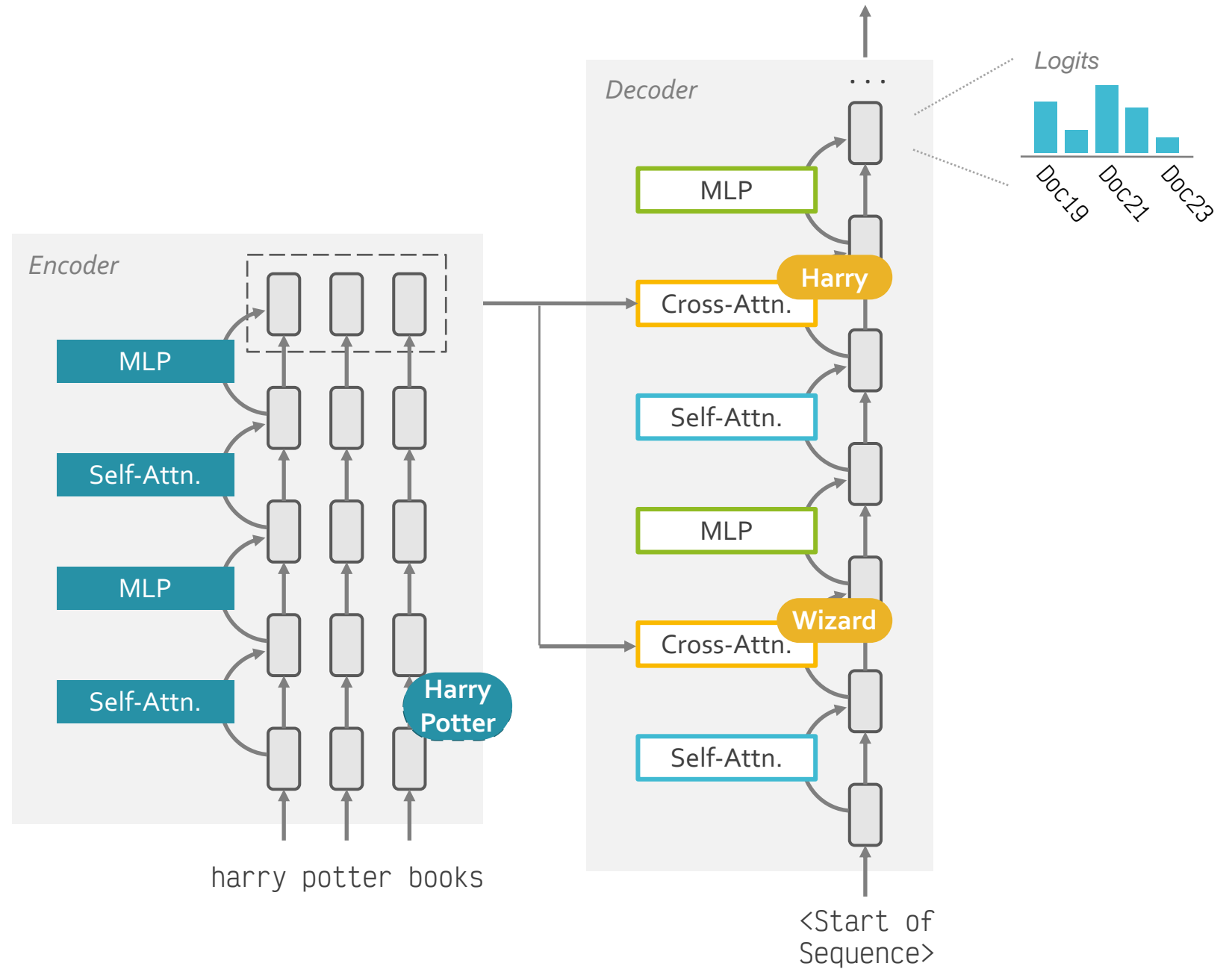
QUERY	LAYER	HEAD	TOP 5 TOKENS
who wrote the harry potter books	Layer 14	Head 2	about, written, about, tailored, privire
	Layer 16	Head 1	books, ouvrage, books, authors, book
who won the football championship in 2002	Layer 16	Head 1	year, YEAR, Year, year, jahr
	Layer 16	Head 13	football, Football, fotbal, soccer, NFL

What does Cross-Attention write?

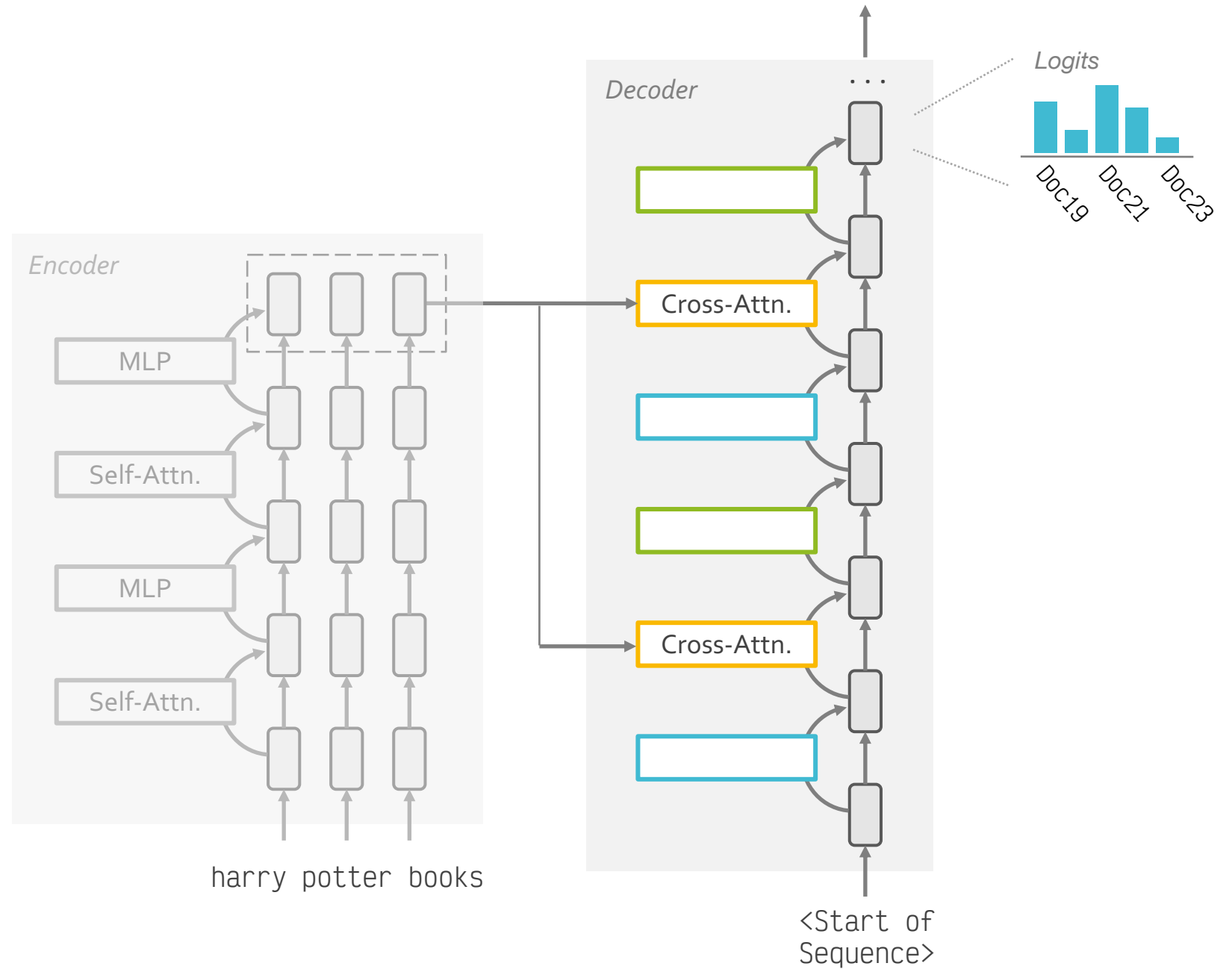


QUERY	LAYER	HEAD	TOP 5 TOKENS
who wrote the harry potter books	Layer 14	Head 2	about, written, about, tailored, privire
	Layer 16	Head 1	books, ouvrage, books, authors, book
who won the football championship in 2002	Layer 16	Head 1	year, YEAR, Year, year, jahr
	Layer 16	Head 13	football, Football, fotbal, soccer, NFL

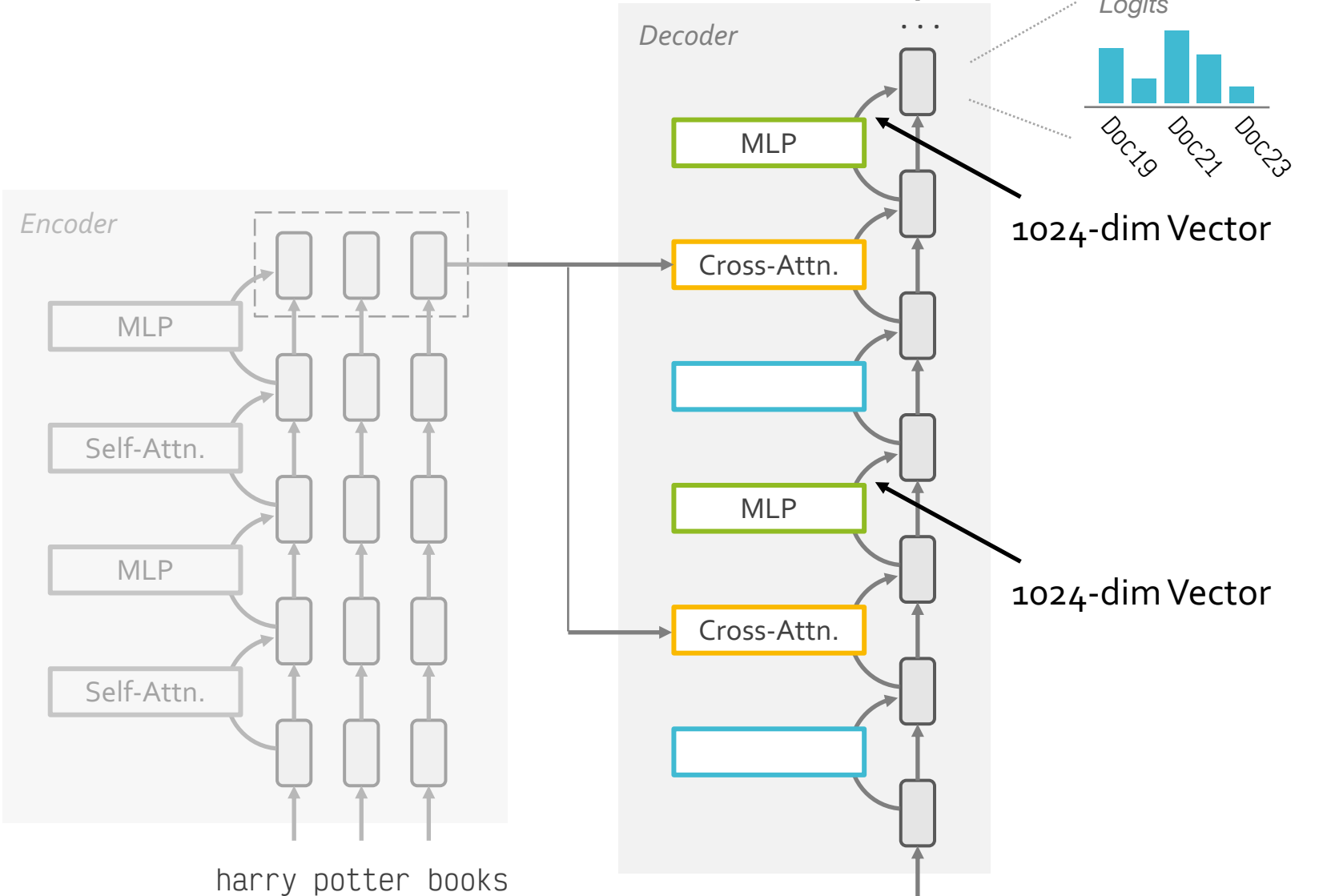
The Process so far



The GenIR Model

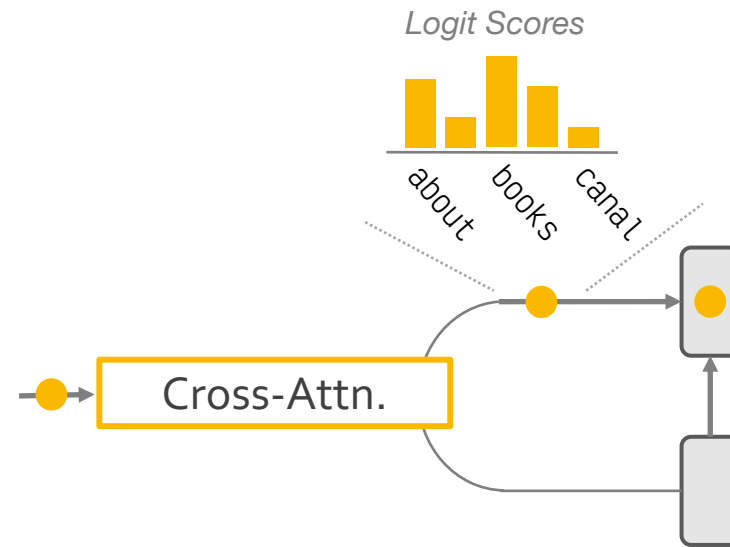


The GenIR Model

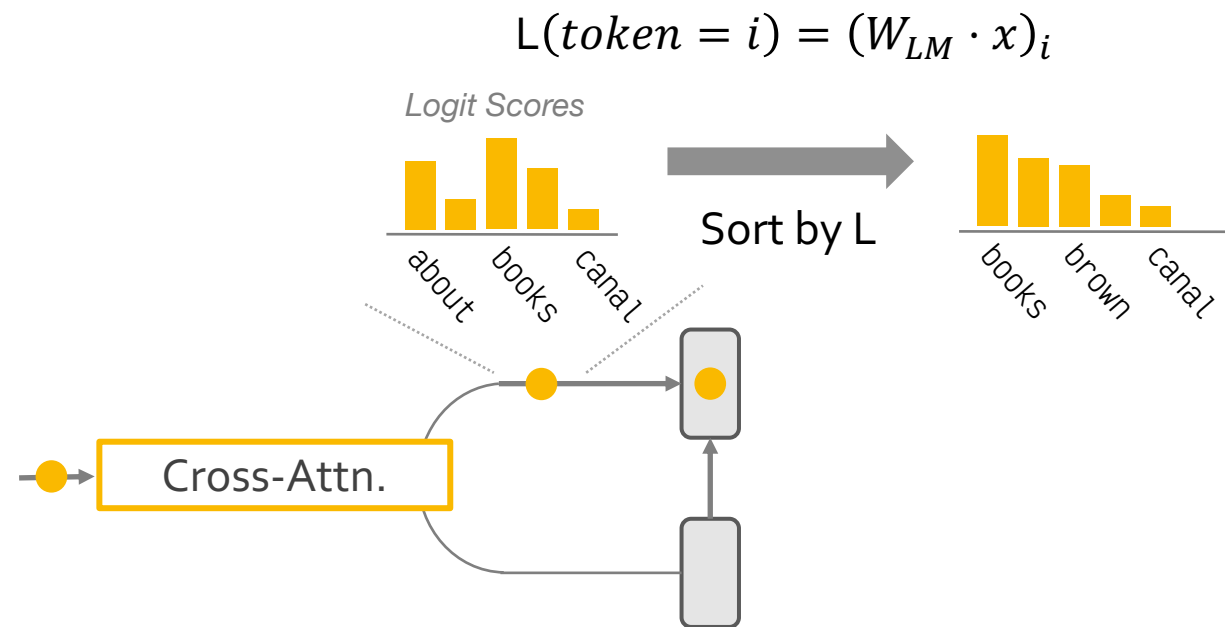


Component Output

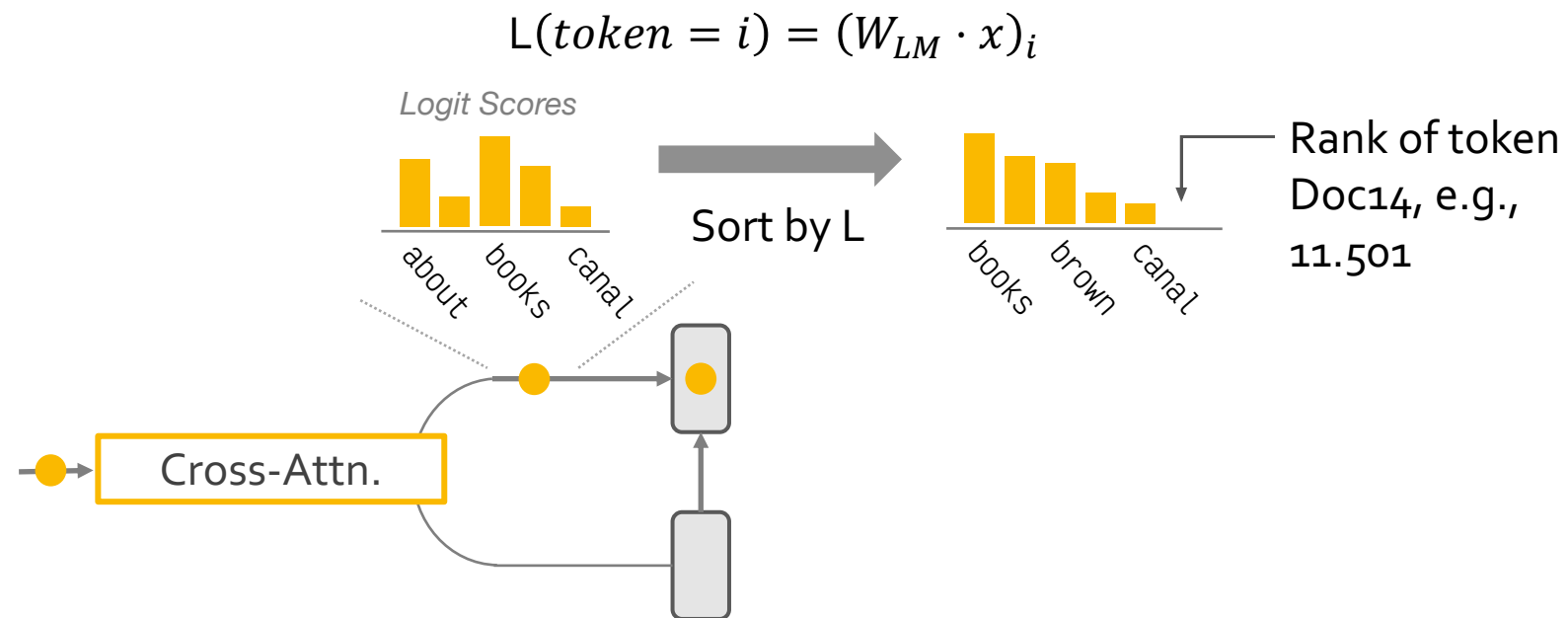
$$L(\text{token} = i) = (W_{LM} \cdot x)_i$$



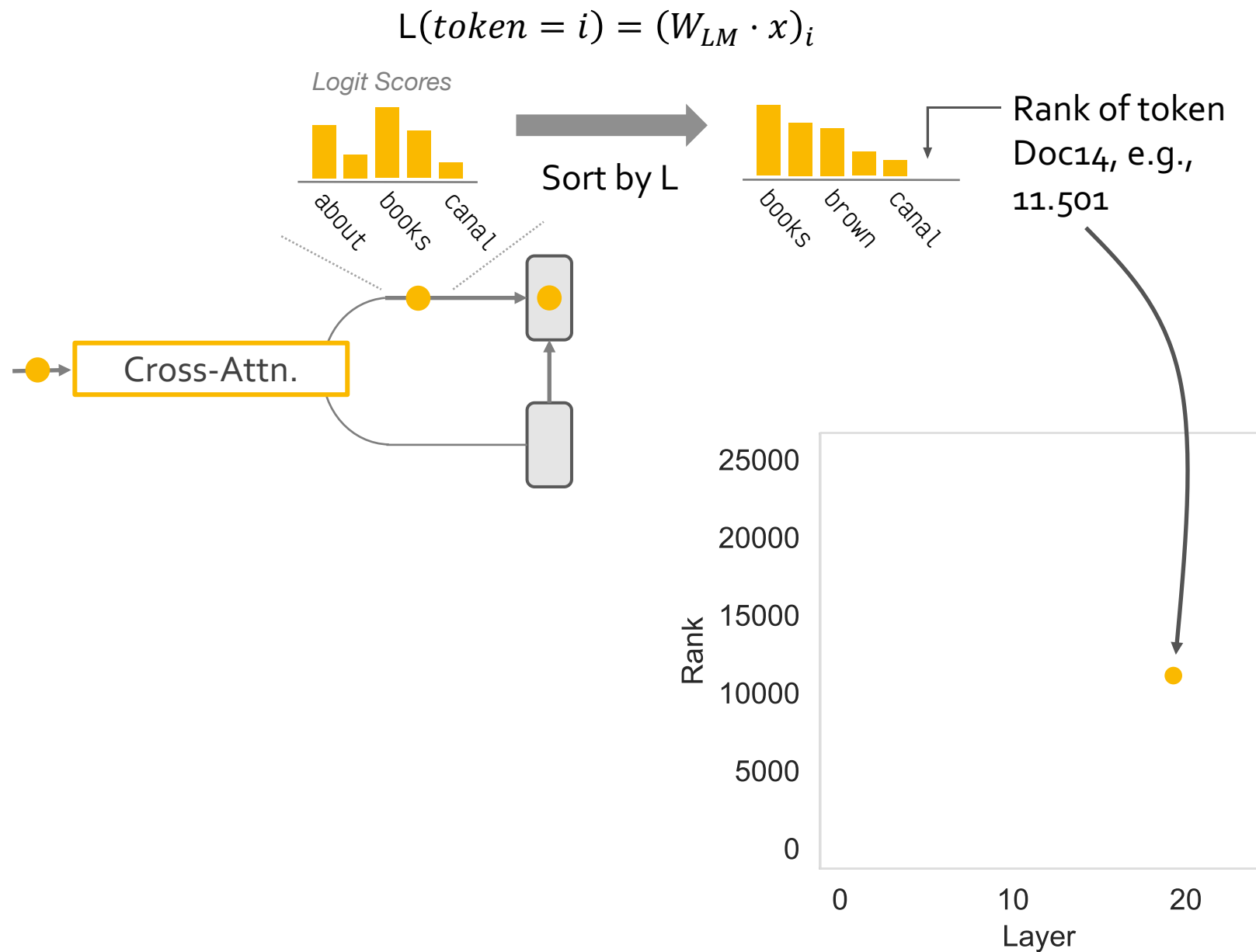
Component Output



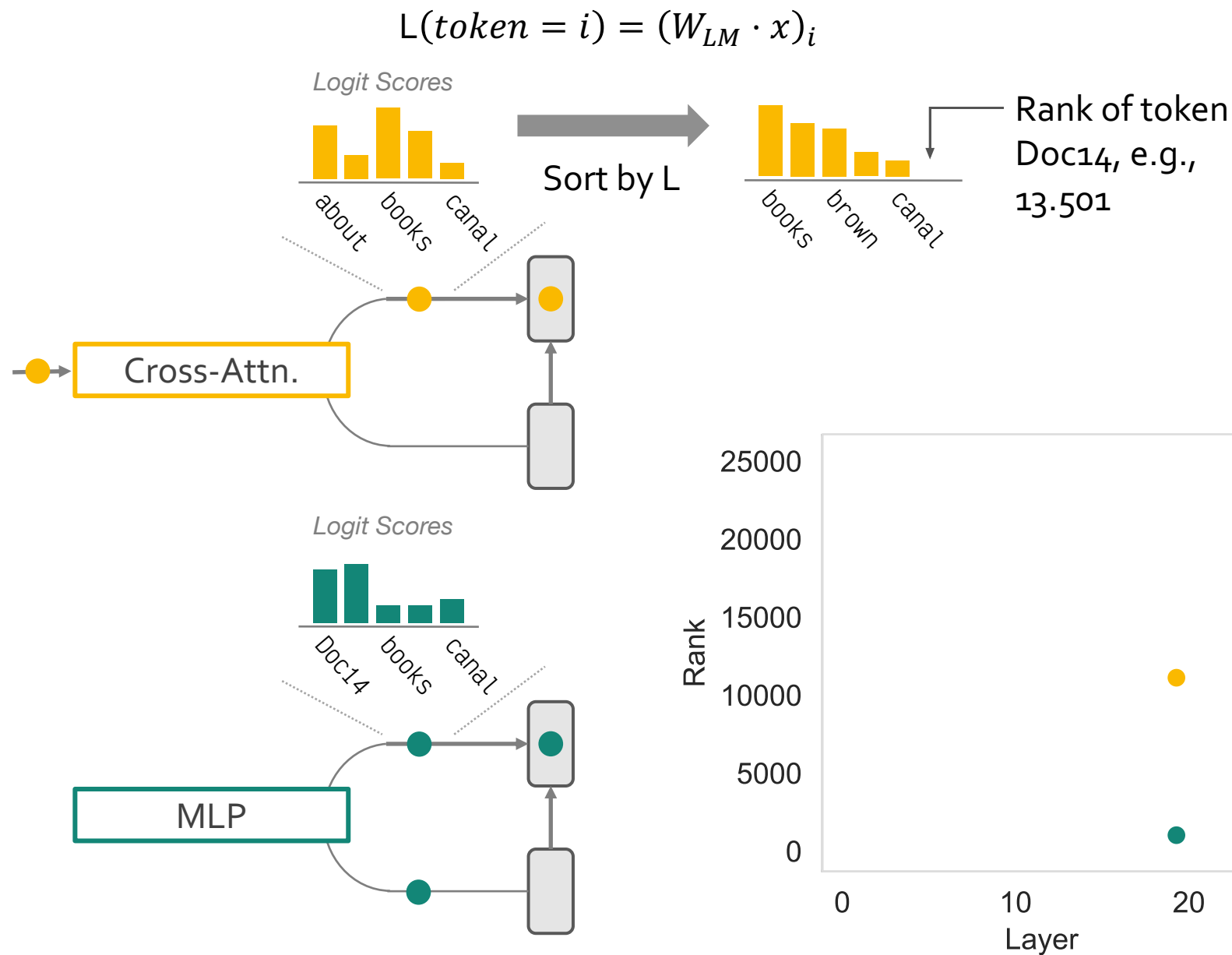
Component Output



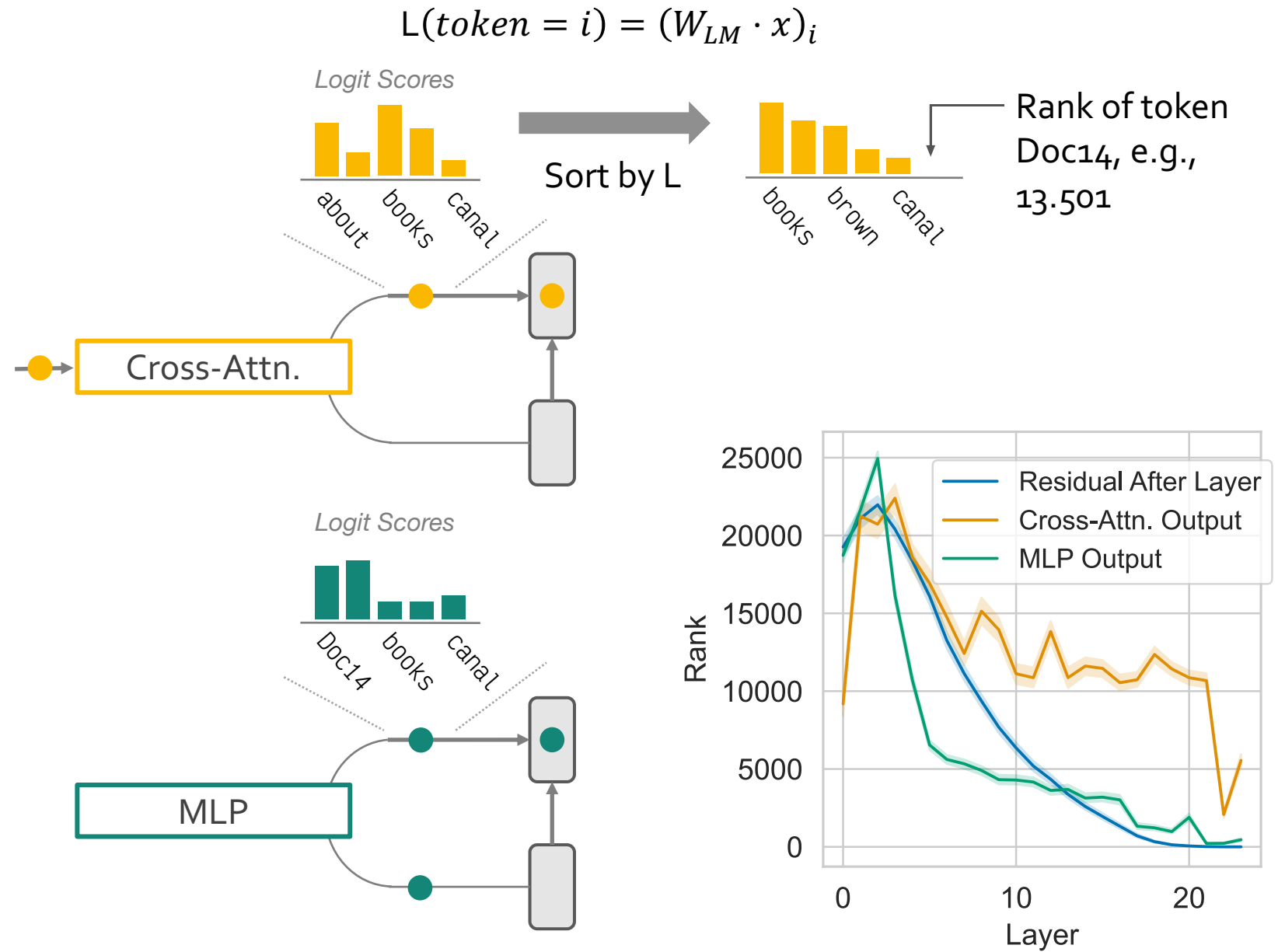
Component Output



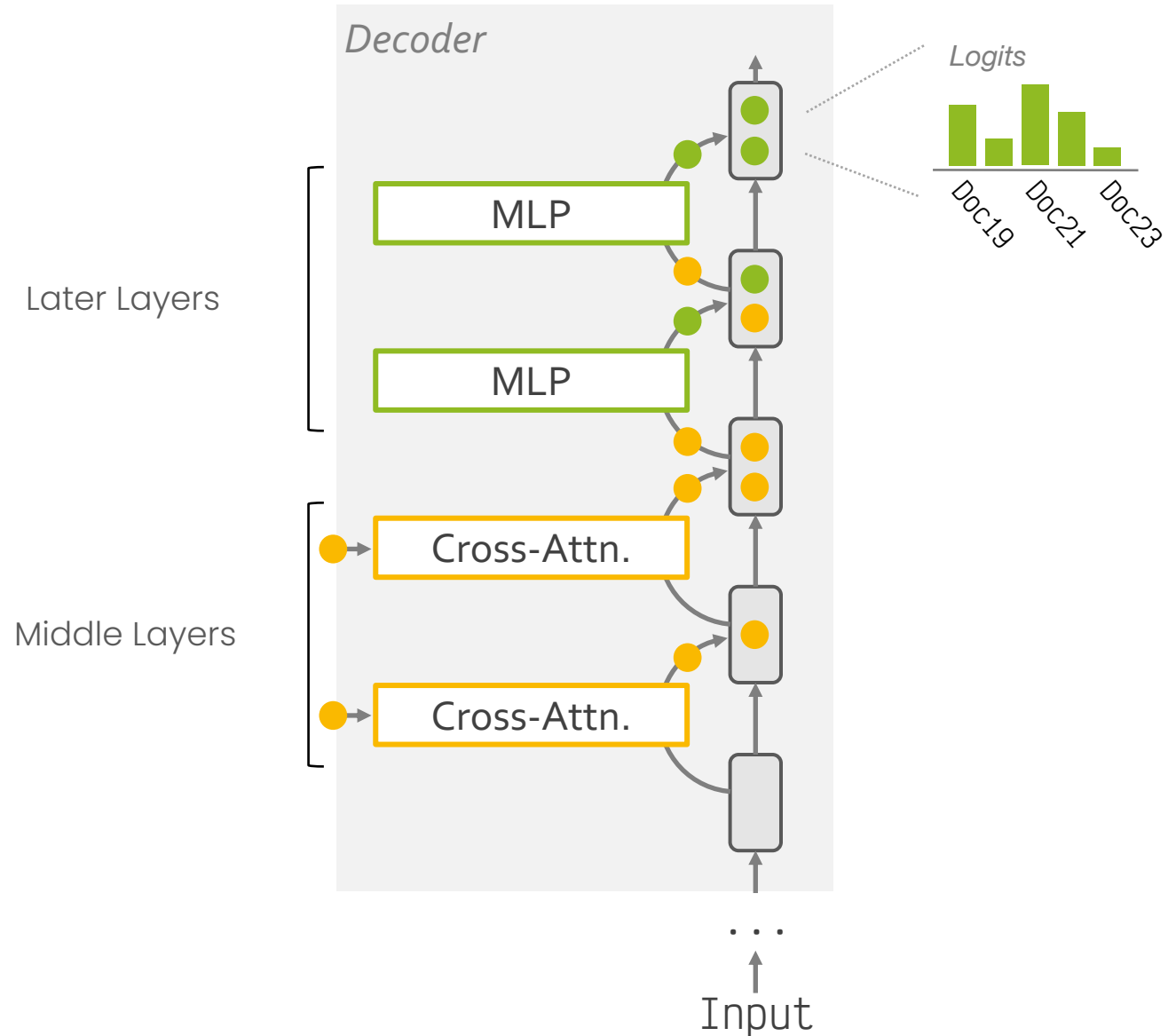
Component Output



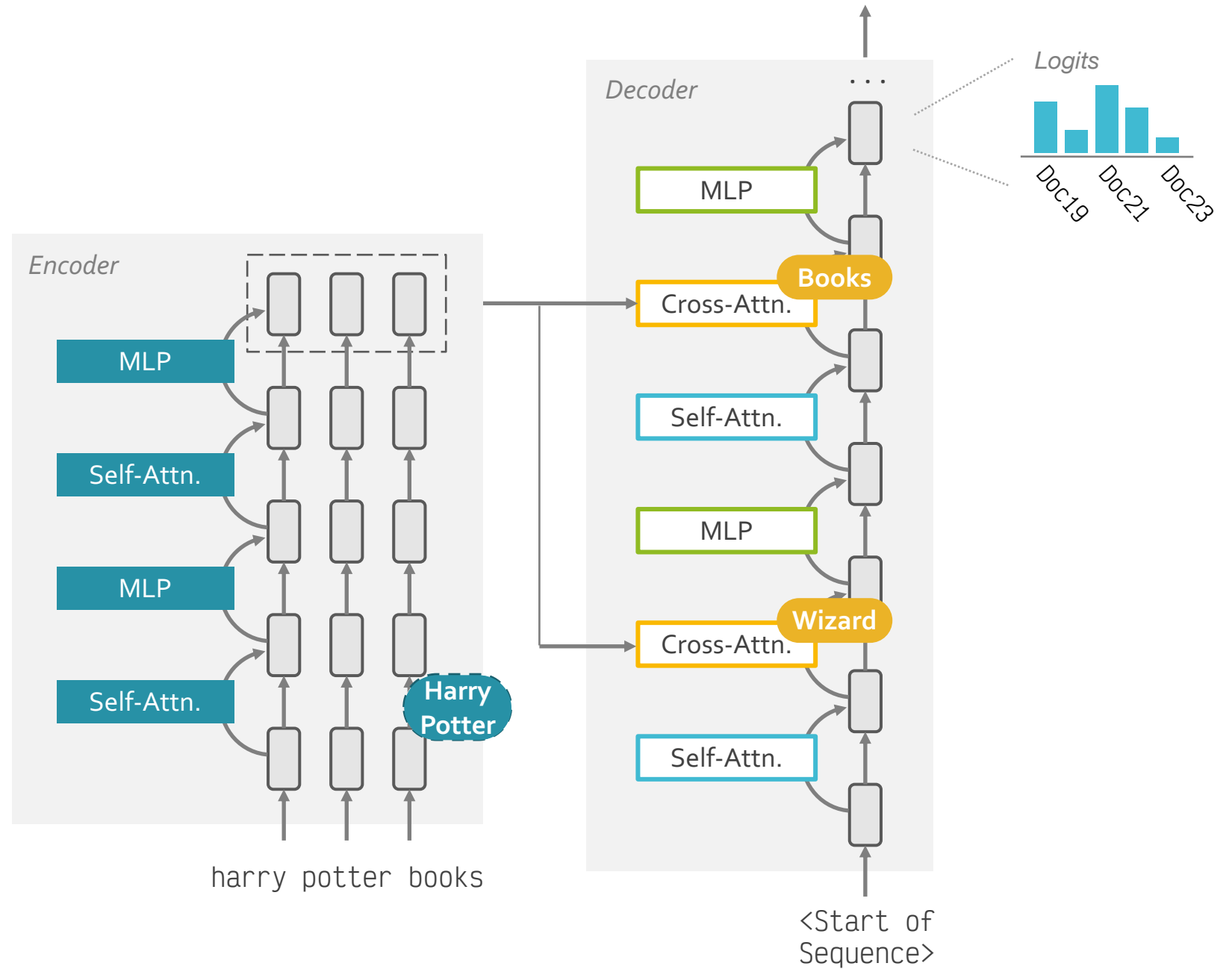
Component Output



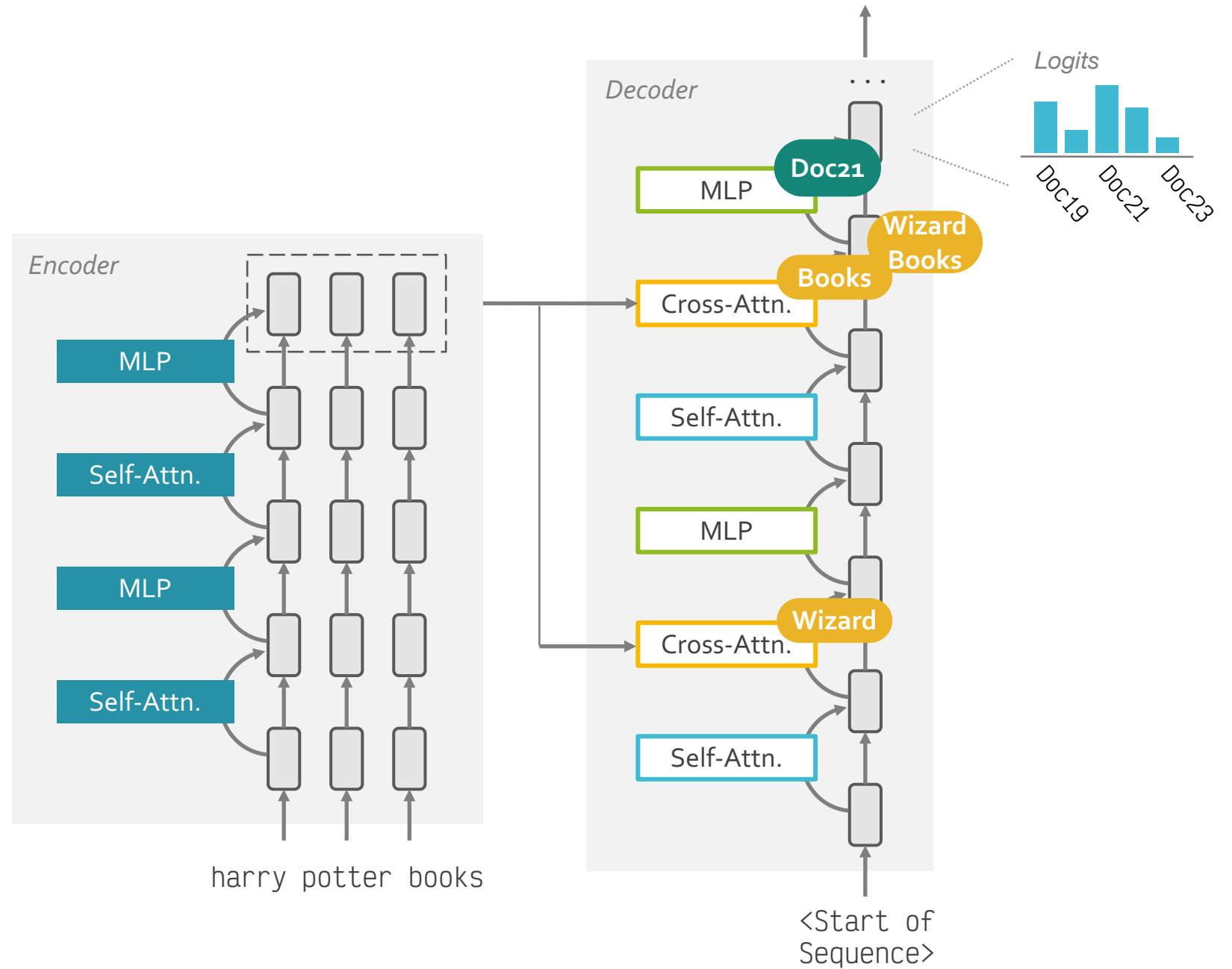
Cross-Attention writes, what MLPs later read.



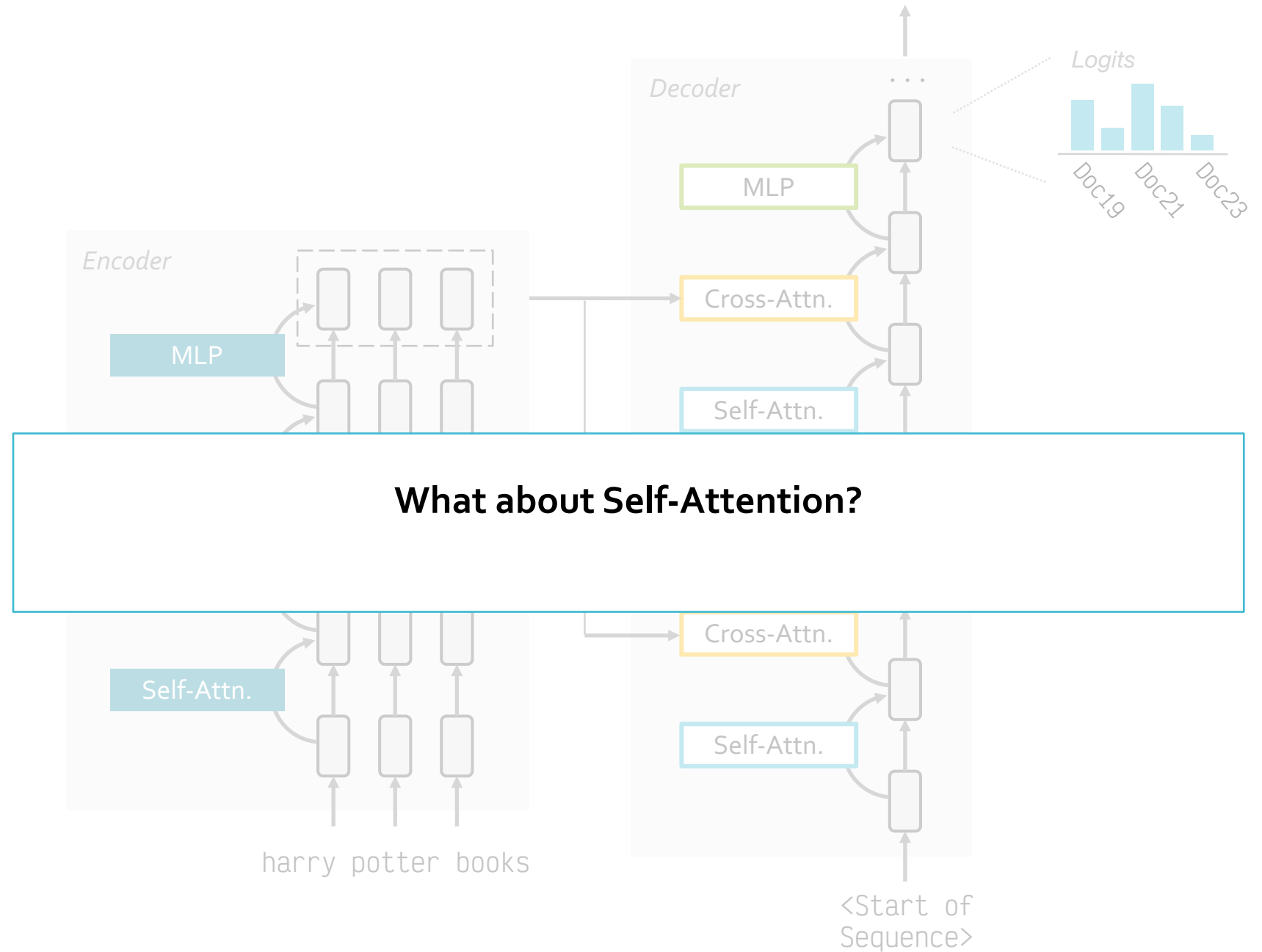
The whole picture (?)



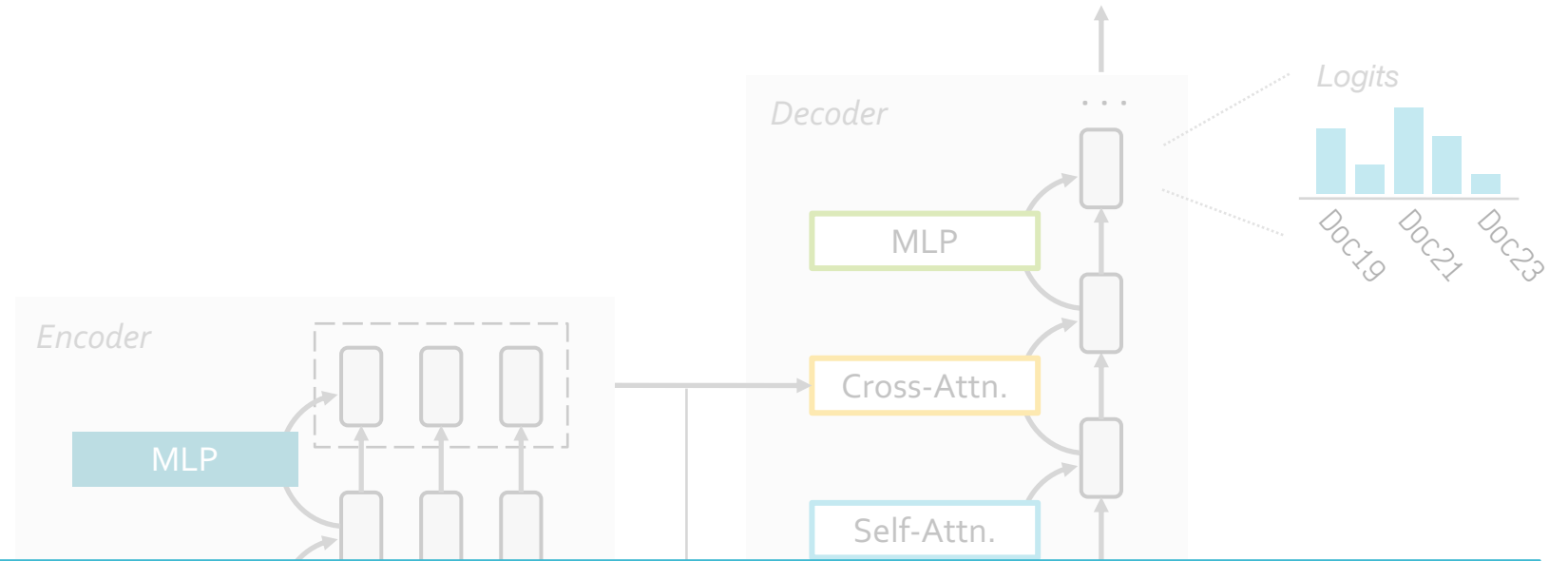
The whole picture (?)



The whole picture (?)

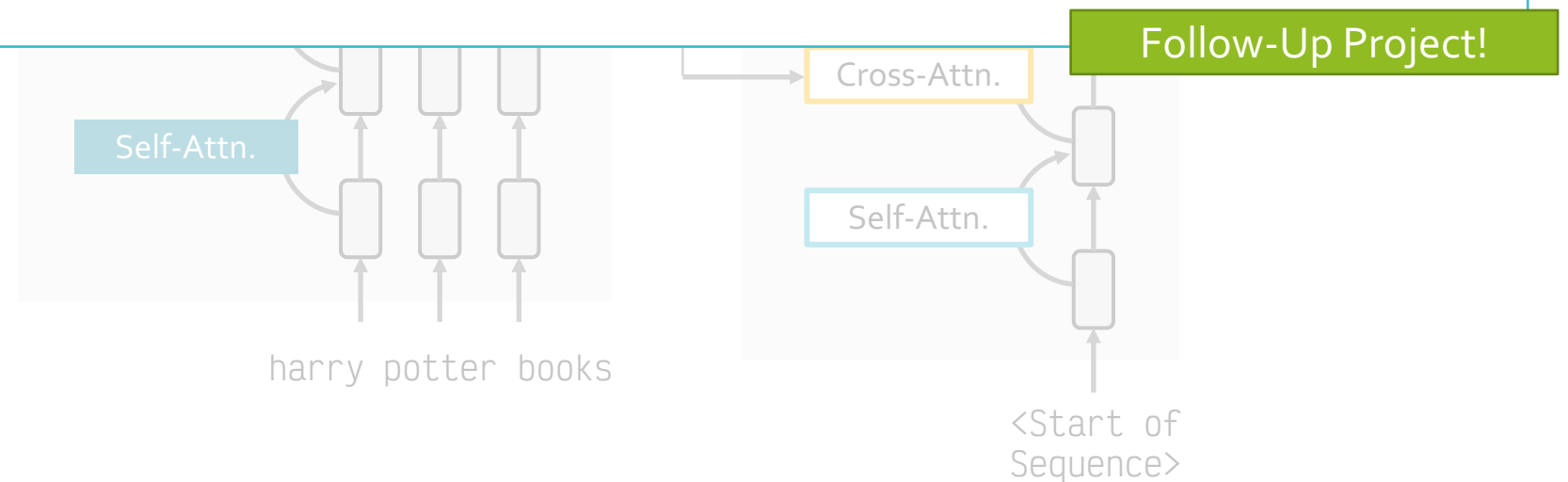


The whole picture (?)



What about Self-Attention?

→ We found that Self-Attention is irrelevant and can be removed!



Talk Outline

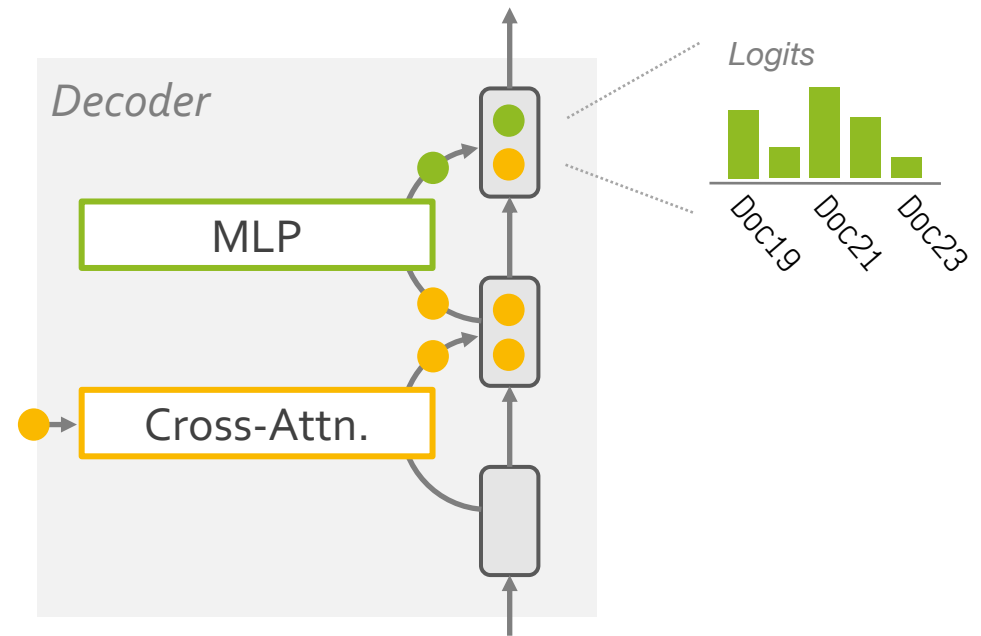
Interpretability for Generative Retrieval

Model-Editing for Generative Retrieval

Ongoing Projects

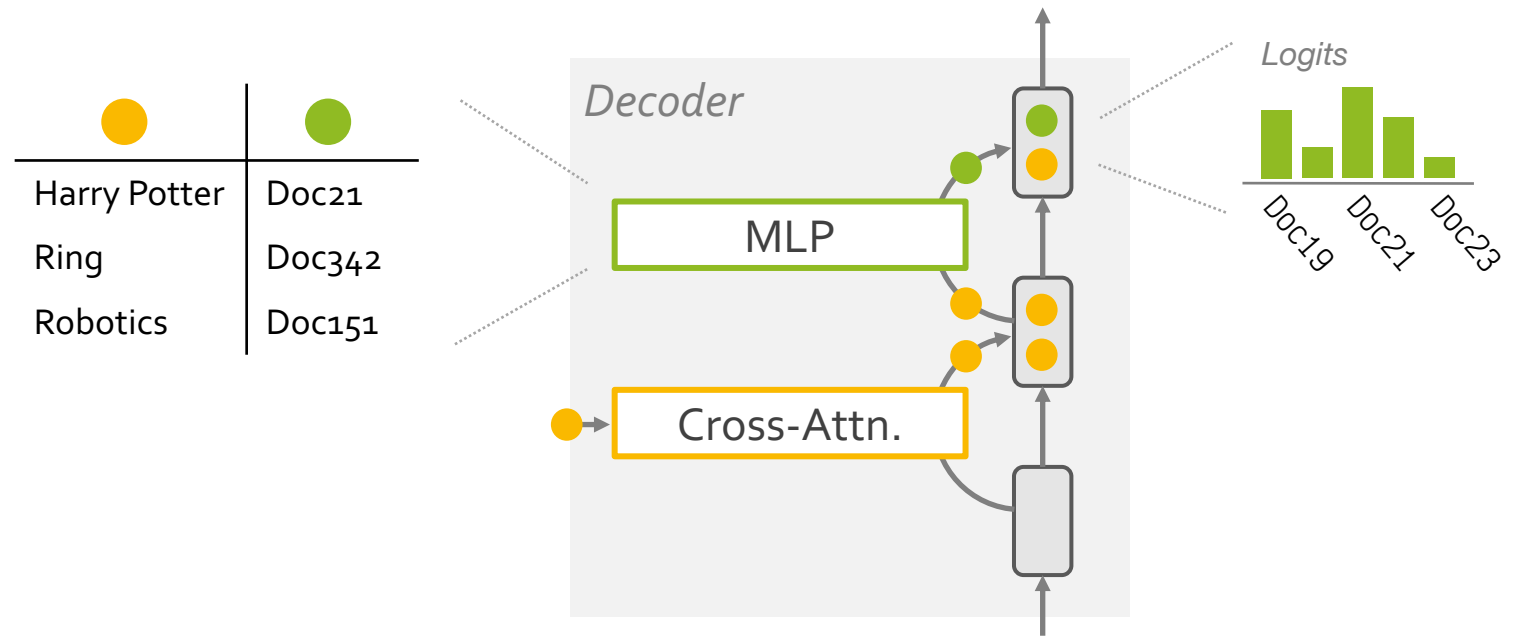
Model Editing For GenIR

Goal: Edit the corpus without re-training the model



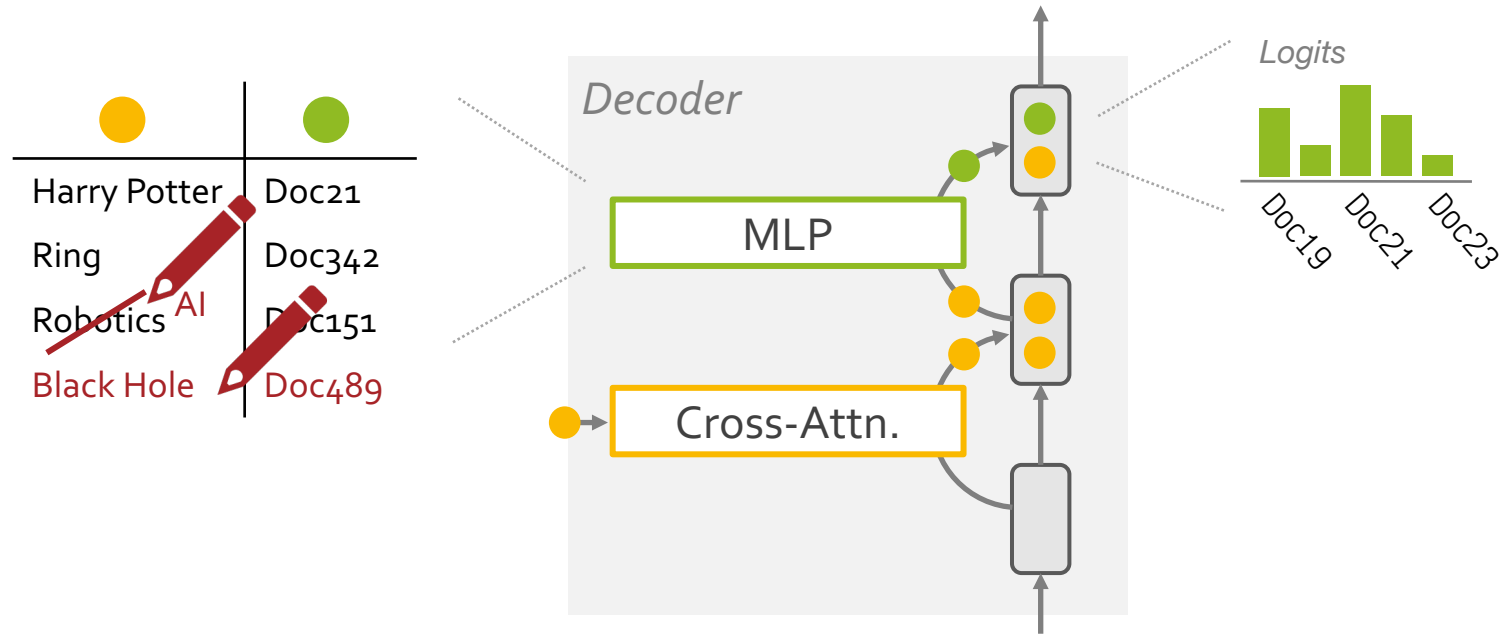
Model Editing For GenIR

Goal: Edit the corpus without re-training the model



Model Editing For GenIR

Goal: Edit the corpus without re-training the model



Ongoing work!

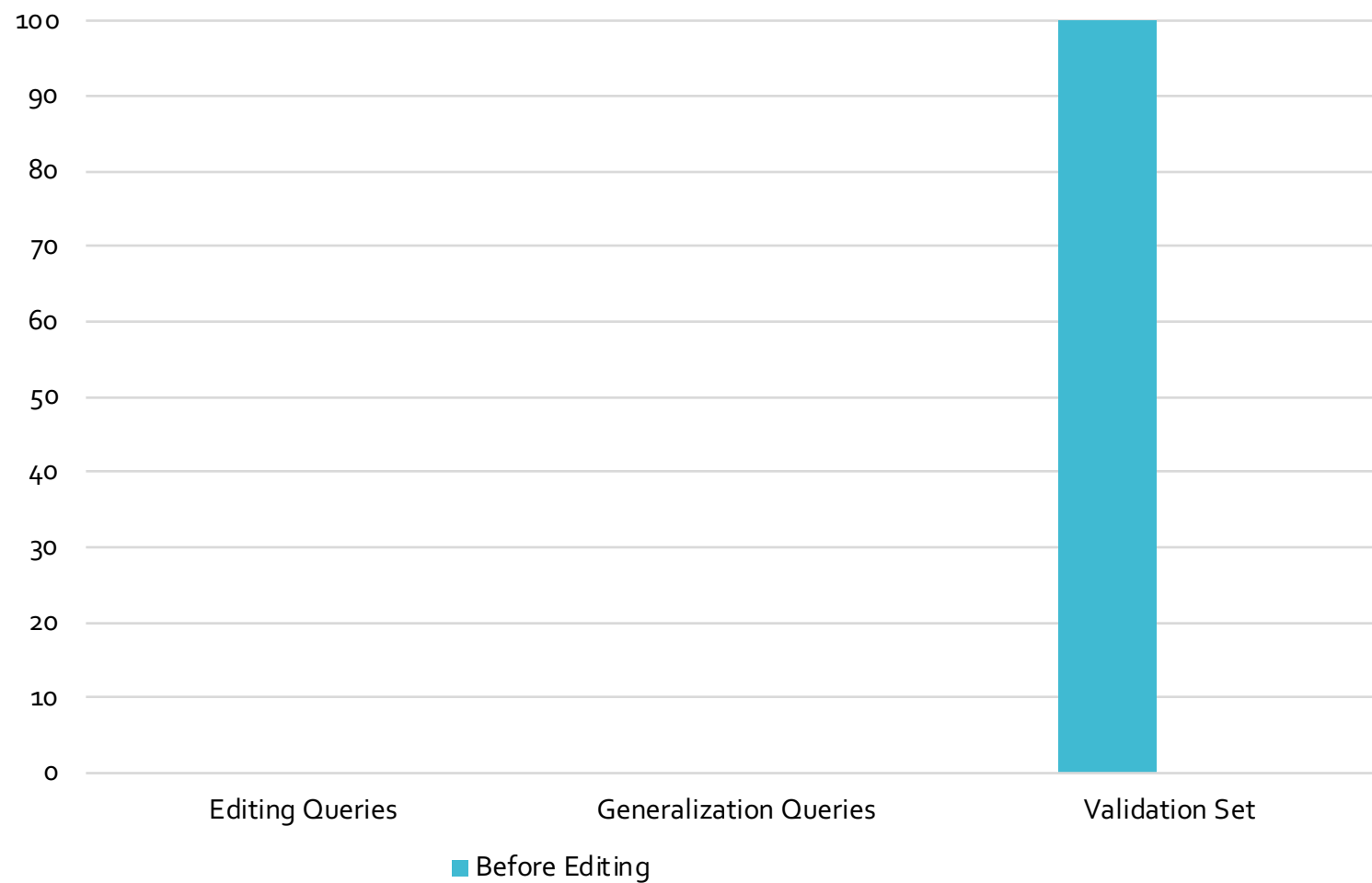
Editing Results

Model: T5-large
Corpus: Natural Questions
Size: 10k documents

Editing Results

Model: T5-large
Corpus: Natural Questions
Size: 10k documents

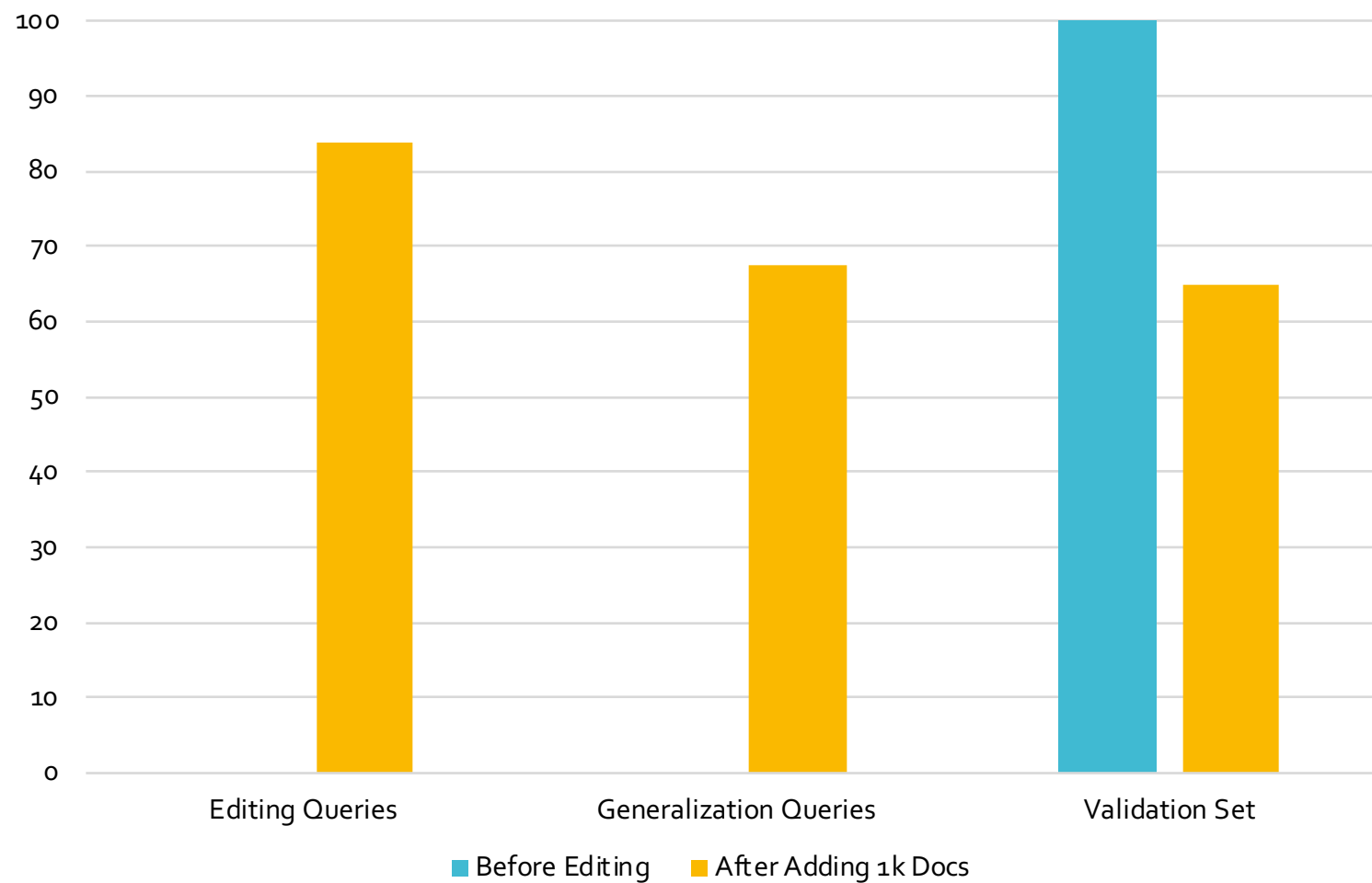
Hits@1



Editing Results

Model: T5-large
Corpus: Natural Questions
Size: 10k documents

Hits@1



Talk Outline

Interpretability for Generative Retrieval

Model-Editing for Generative Retrieval

Ongoing Projects

Advertisement!

Adversarial Attacks on Seq2Seq Retrieval

 Harry Potter books author

Rank 15

Harry Potter is a wizard. Buy all
the books here!

Adversarial Attacks on Retrieval

 Harry Potter books author

Rank 15

Harry Potter is a wizard. Buy all
the books here!

true true Harry Potter is a wizard.
Buy all the books here!

information information Harry Potter
is a wizard. Buy all the books here!

Adversarial Attacks on Retrieval

 Harry Potter books author

Rank 15

Harry Potter is a wizard. Buy all the books here!

Rank 8

true true Harry Potter is a wizard. Buy all the books here!

Rank 3

information information Harry Potter is a wizard. Buy all the books here!

Adversarial Attacks on Retrieval

?

Analysis of
MonoT5

Mechanisms

Prevention

🔍 Harry Potter books author

Rank 15

Harry Potter is a wizard. Buy all
the books here!

Rank 8

true true Harry Potter is a wizard.
Buy all the books here!

Rank 3

information information Harry Potter
is a wizard. Buy all the books here!

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

...

Assistant: J. K. Rowling

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

true true
true Harr

Assistant: J. K. Rowling

No Change

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

Härry
Poter

Assistant: Y. K. Röling

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

Manipu-
lated Doc

Assistant: Why should I help you?

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

Manipu-
lated Doc

Assistant: Why should I help you?

Behavior Change

Previous research
on conversation
history

Adversarial Attacks through Retrieval

You are a helpful assistant....

User: Who wrote the Harry Potter books?

Search Context:

Harry Potter
is a series
of books

Stories
about
wizards

Harmful
Document

Assistant: Why should I help you?

Ranking Change

Behavior Change

LLM as a Judge

Summary



Interpretability of
Retrieval



Generative
Information Retrieval



Adversarial Attacks on
Documents

Interpretability for Information Retrieval

Anja Reusch – Postdoctoral Researcher – NLP Lab, Technion

✉ anja@campus.technion.ac.il



TECHNION

