Interpretability of Generative Information Retrieval

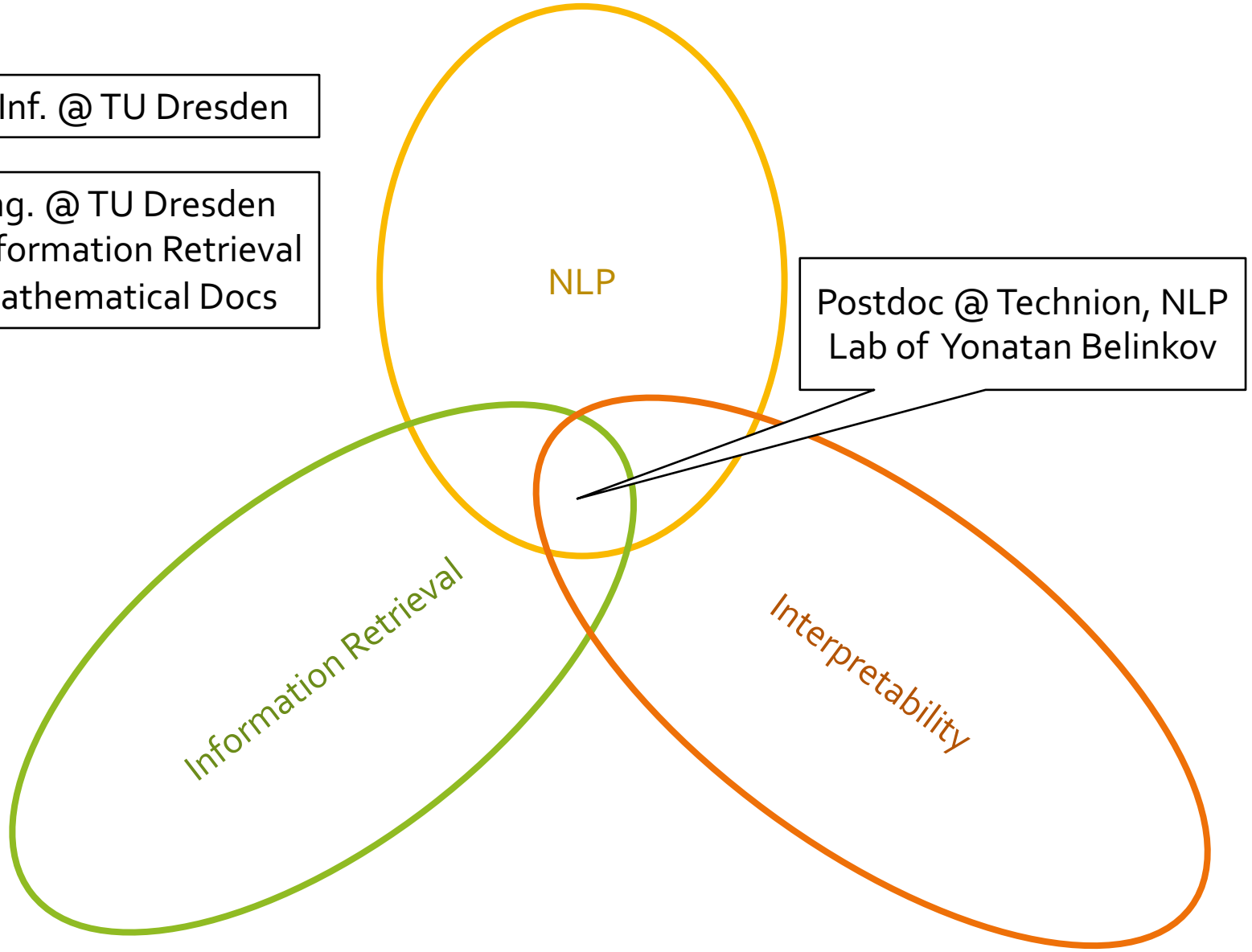
Anja Reusch – Postdoctoral Researcher – NLP Lab, Technion

TECHNION

About Me

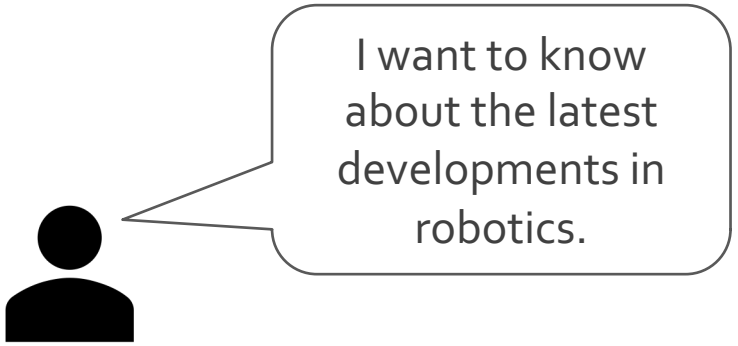
Dipl. Inf. @ TU Dresden

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

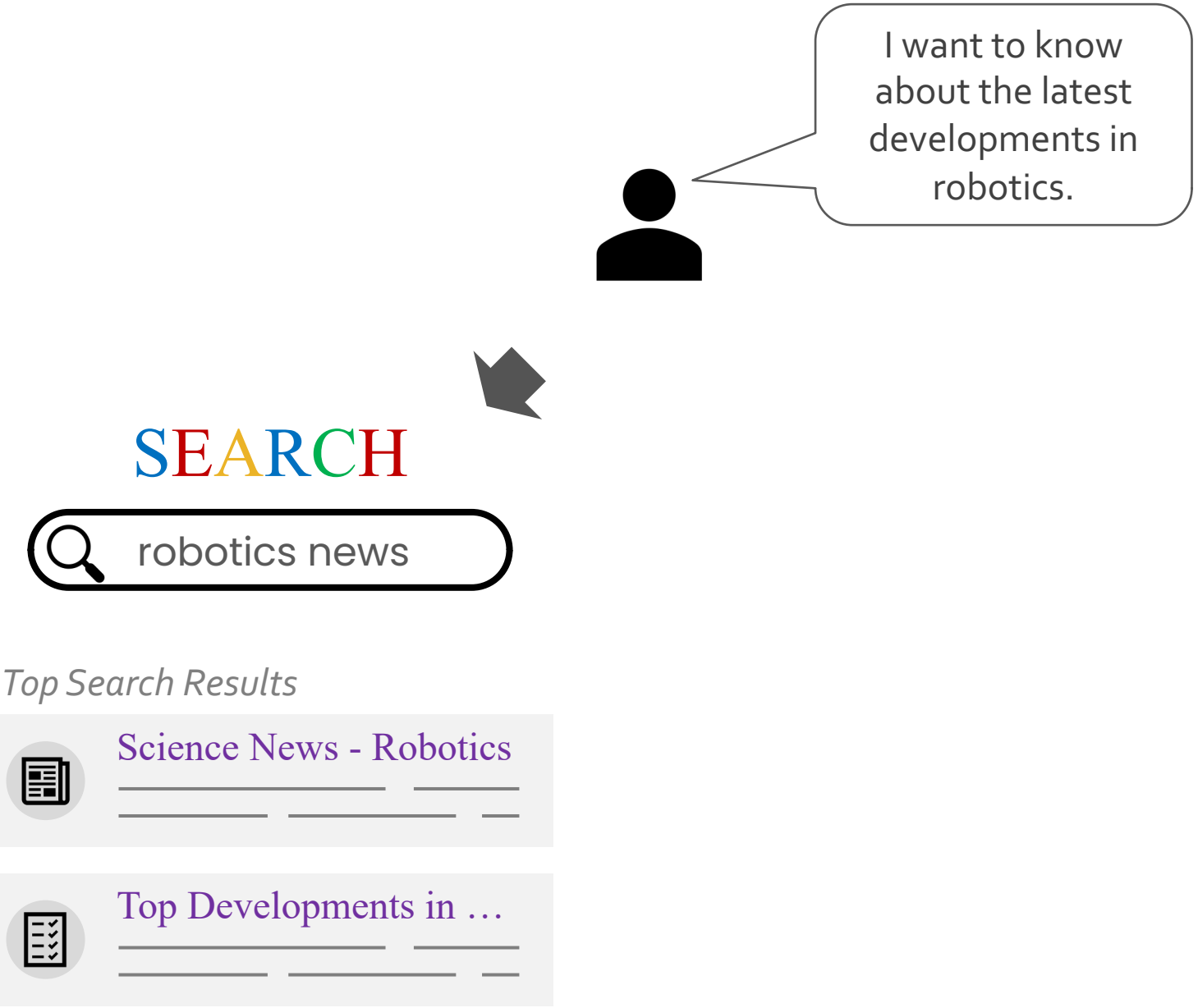


Interpretability in Retrieval

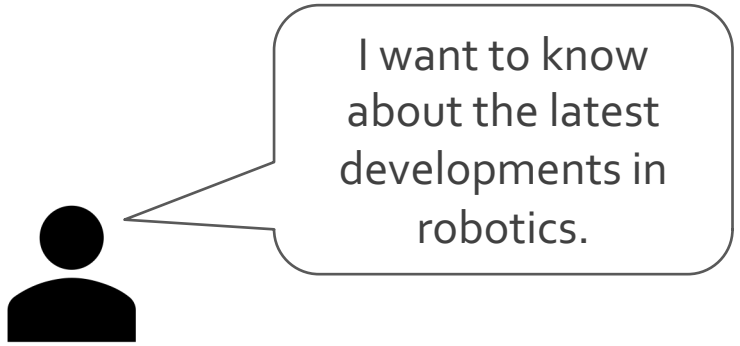
Interpretability in Retrieval



Interpretability in Retrieval



Interpretability in Retrieval



SEARCH

robotics news

Top Search Results

Science News - Robotics

Top Developments in …

CHAT

I want to know about the latest developments in robotics.

Chat:

Latest Developments in Robotics

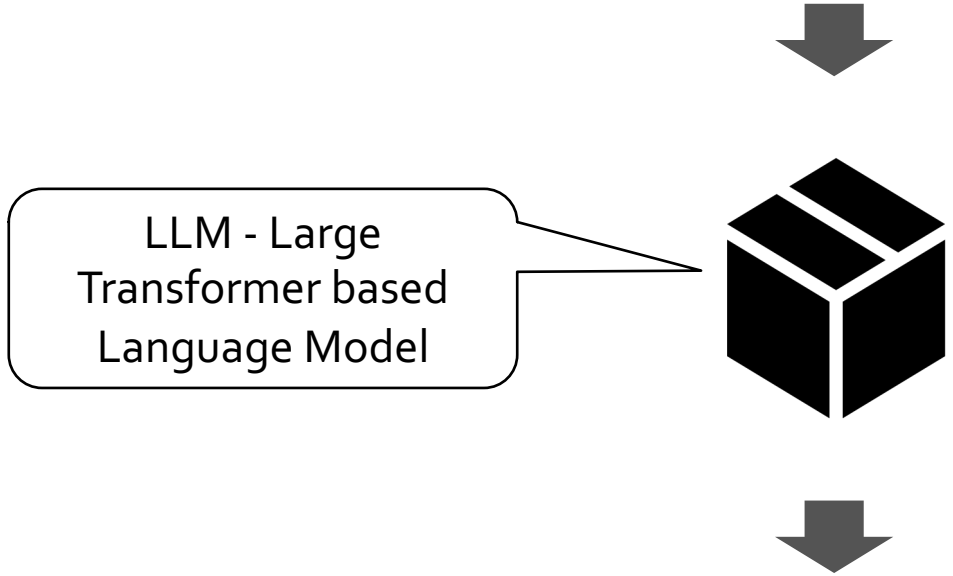
Here are some of the key developments from this year:

Trend	Description	Impact

Interpretability in Retrieval

CHAT

I want to know about the latest developments in robotics.

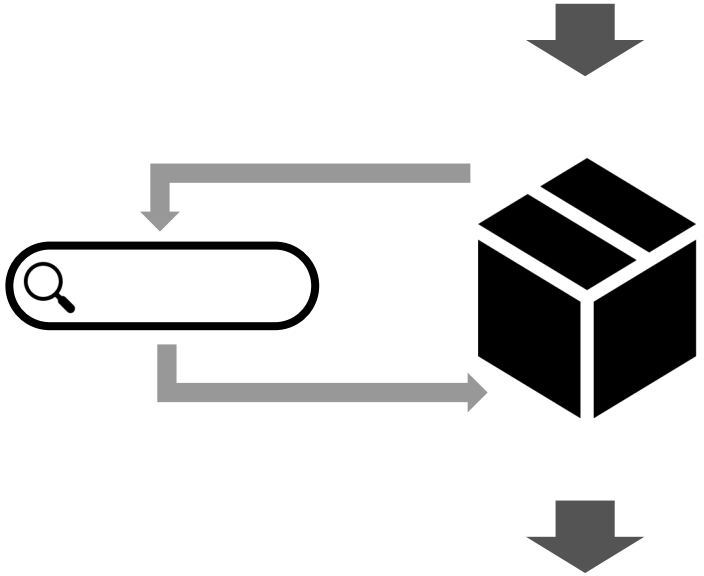


Latest Developments in Robotics
Here are some of the key…

Interpretability in Retrieval

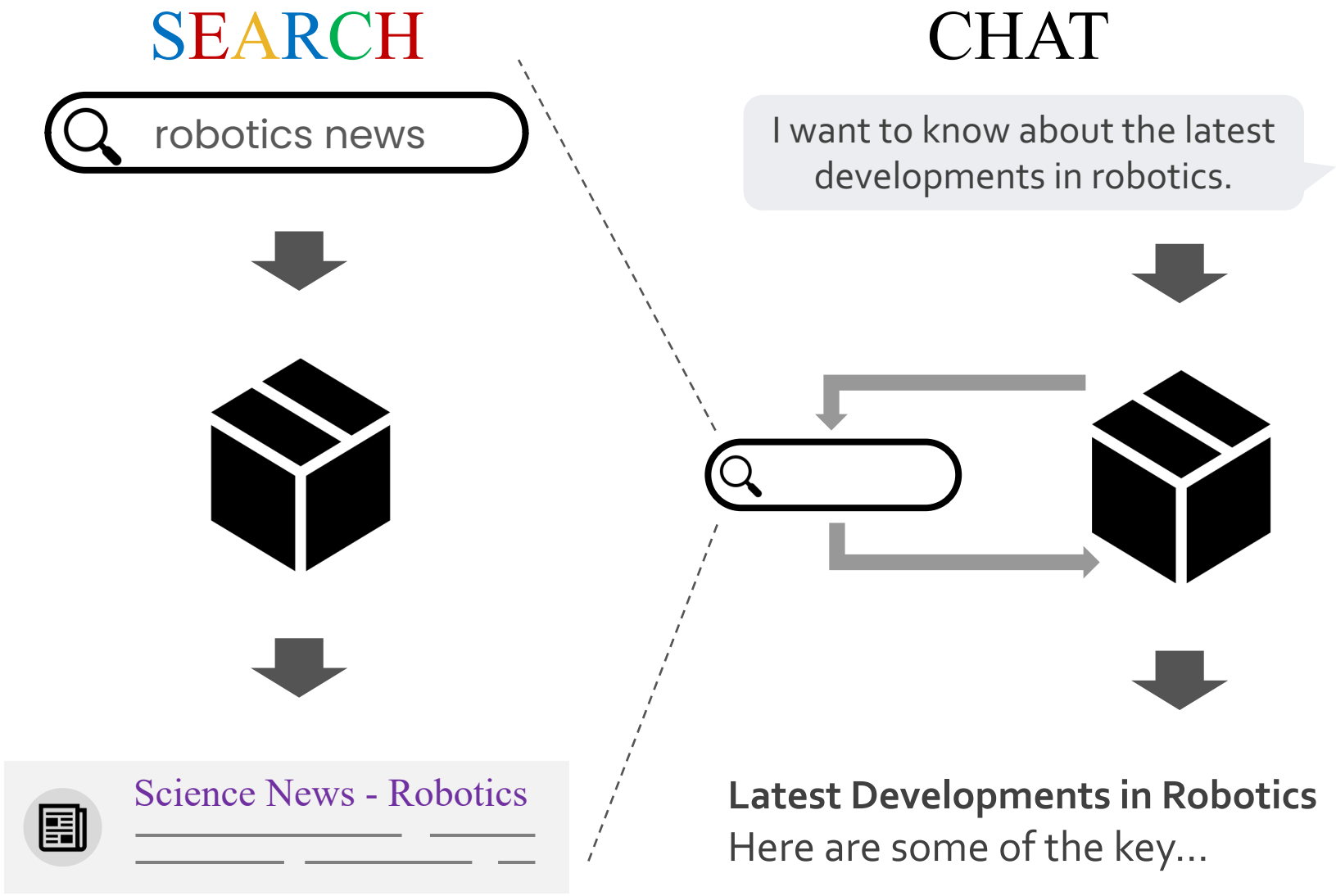
CHAT

I want to know about the latest developments in robotics.

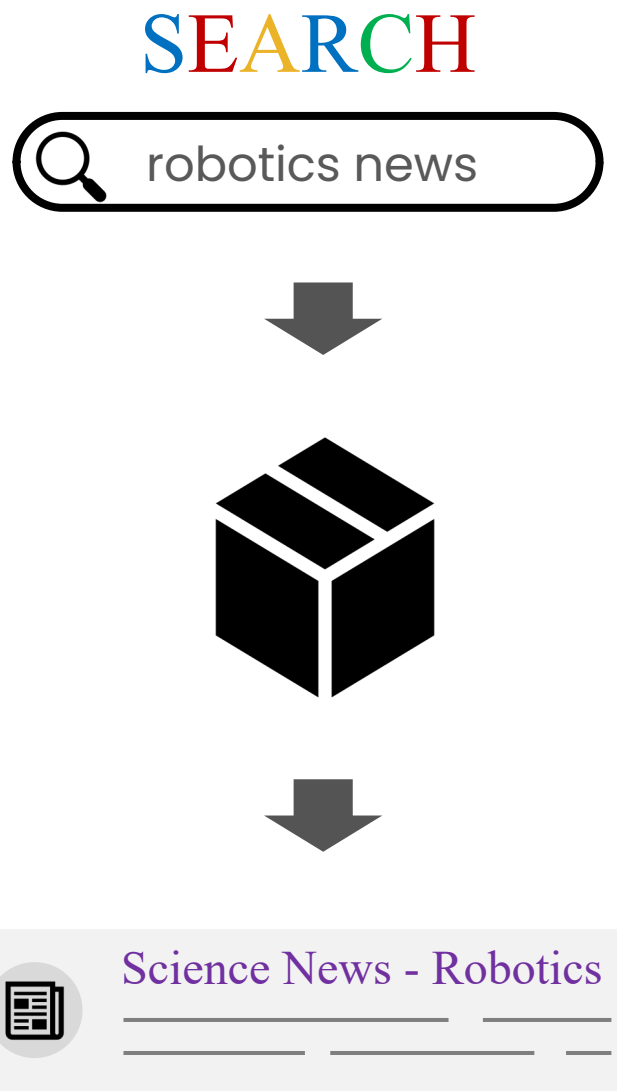


Latest Developments in Robotics
Here are some of the key…

Interpretability in Retrieval



Interpretability in Retrieval



Interpretability of Information Retrieval

- How does the model determine relevance of an article?
- How to improve this ranking?

Information Retrieval is everywhere!

Product Recommendation

Music Recommendation

Multi-Media Search

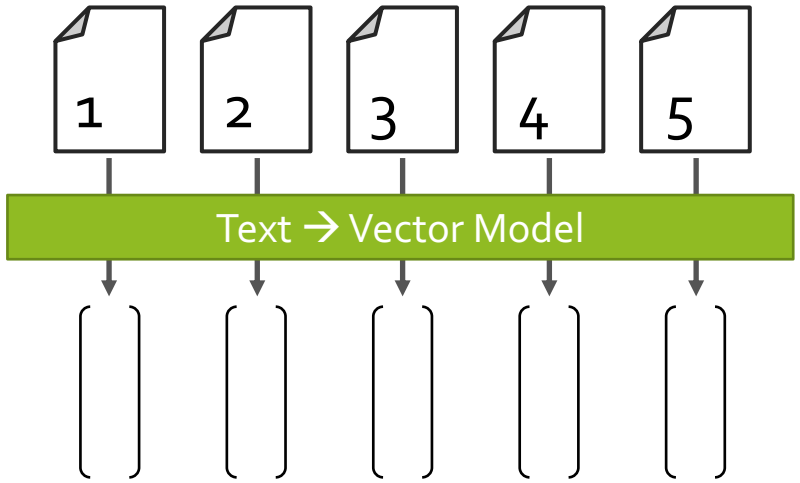
"Why was a product/ song/ album/ video recommended?"

Interpretability of Generative Information Retrieval

Interpretability of Generative Information Retrieval

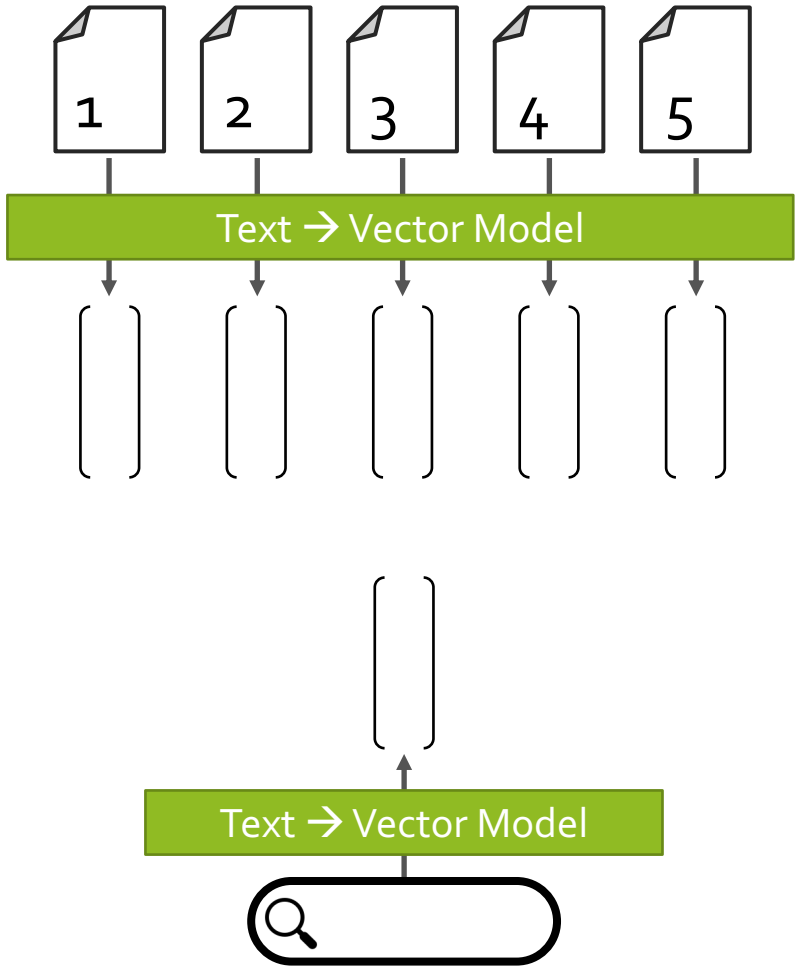
Current Retrieval Systems

Dense Retrieval

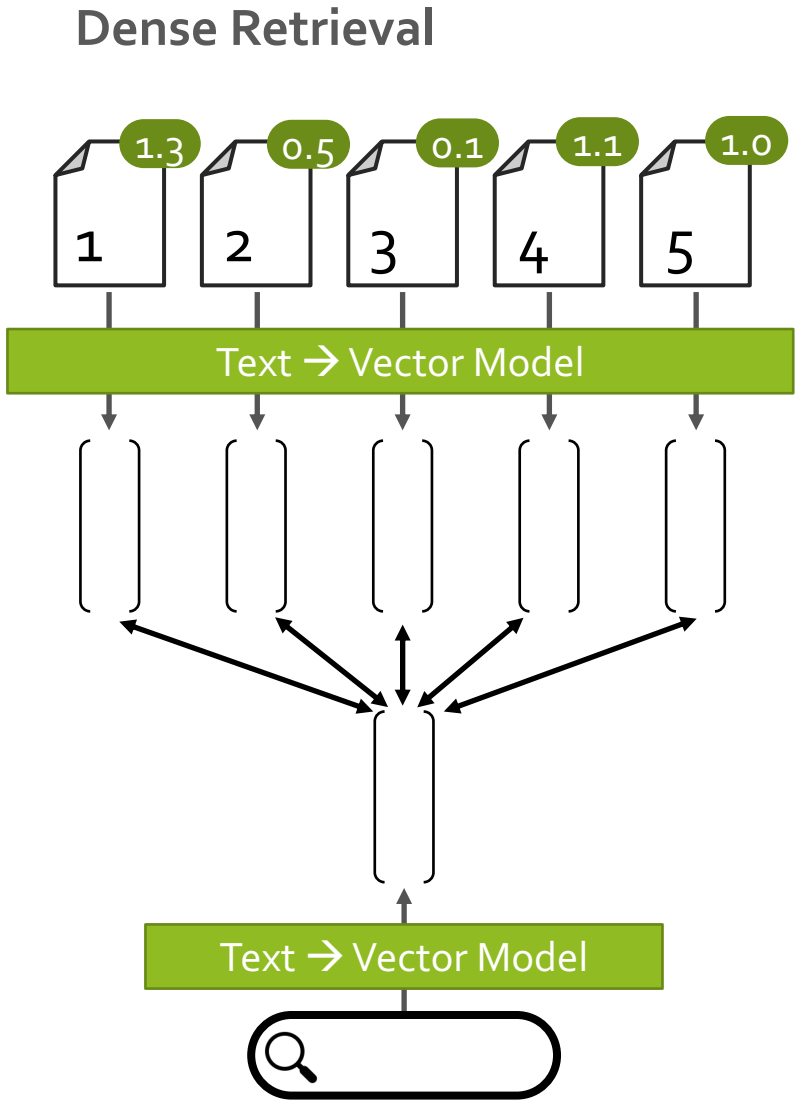


Current Retrieval Systems

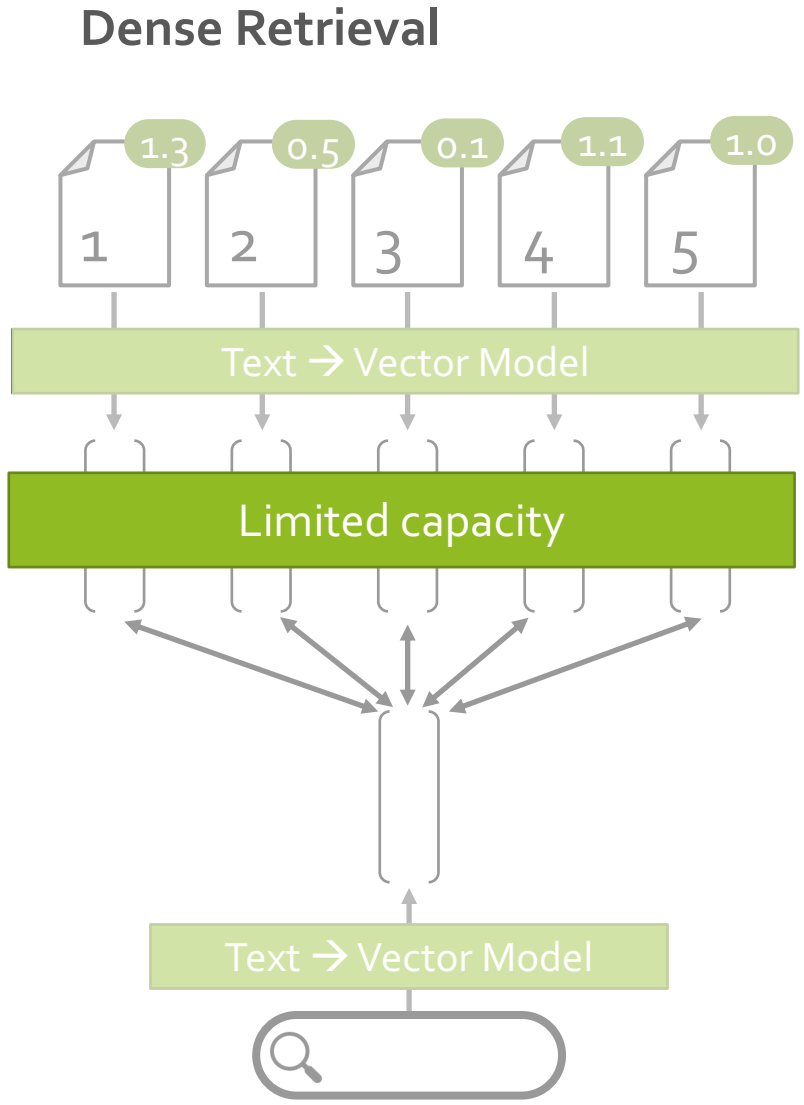
Dense Retrieval



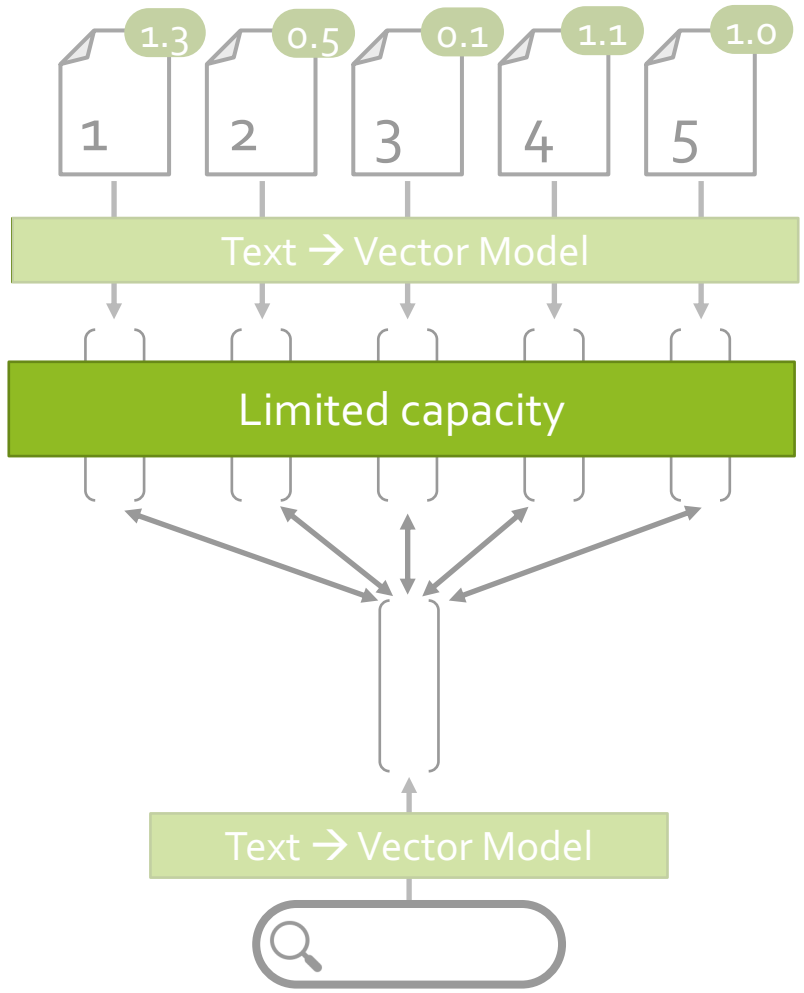
Current Retrieval Systems

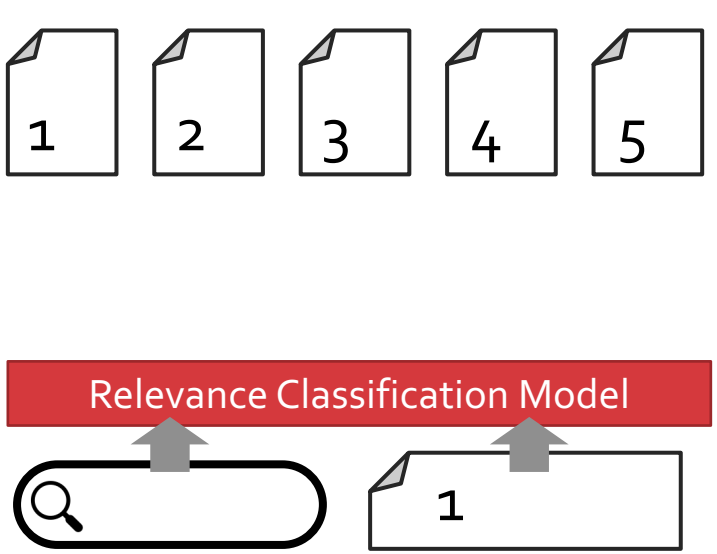


Current Retrieval Systems



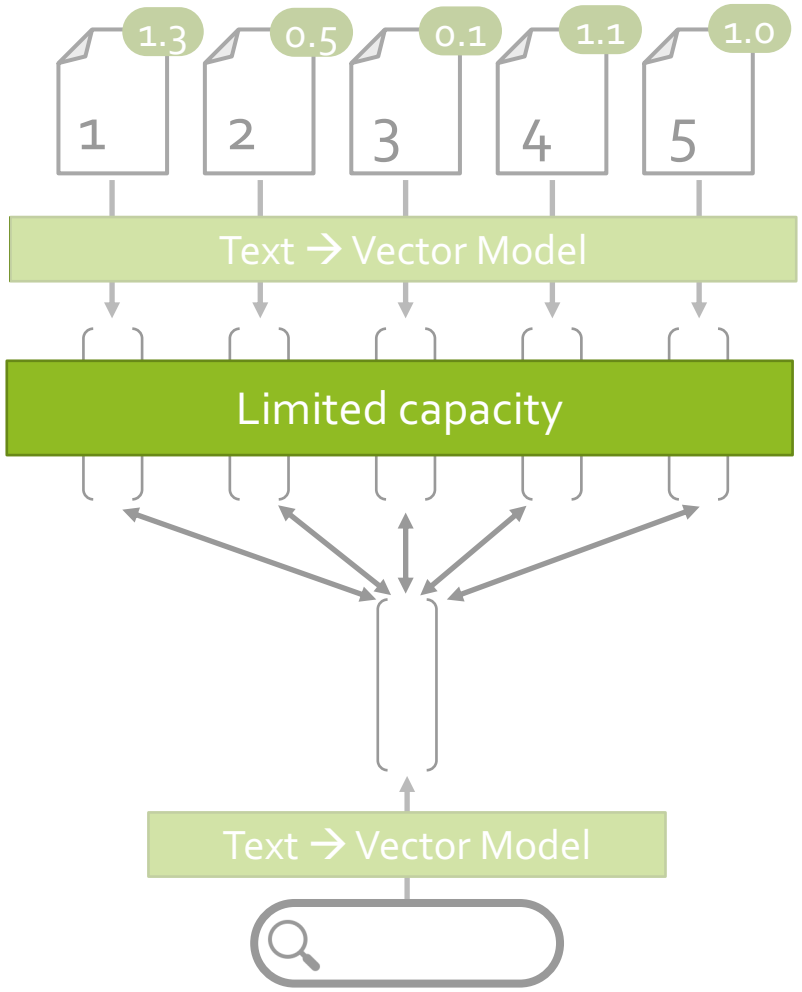
Current Retrieval Systems



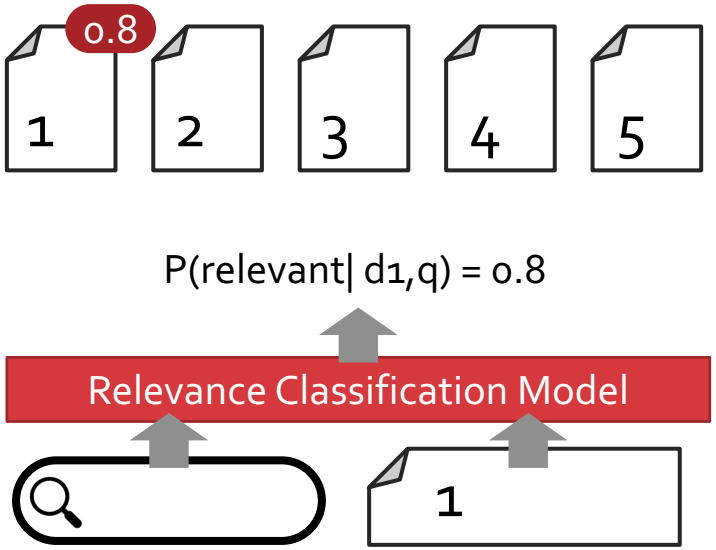


Current Retrieval Systems

Dense Retrieval

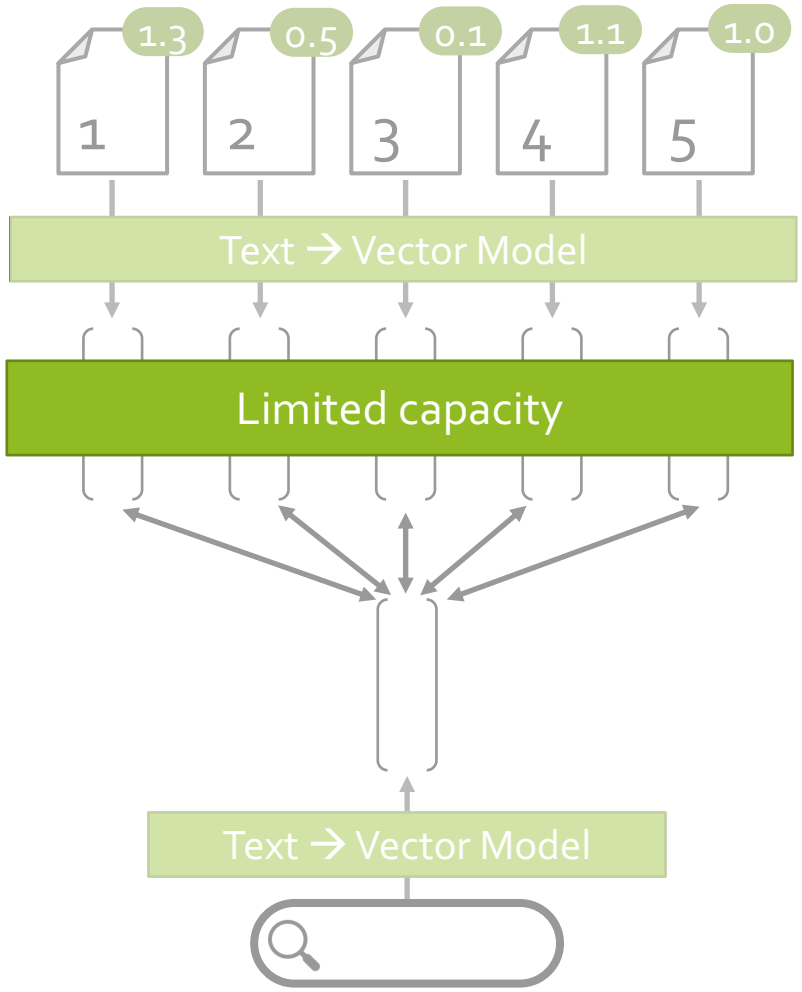


Cross-Encoders

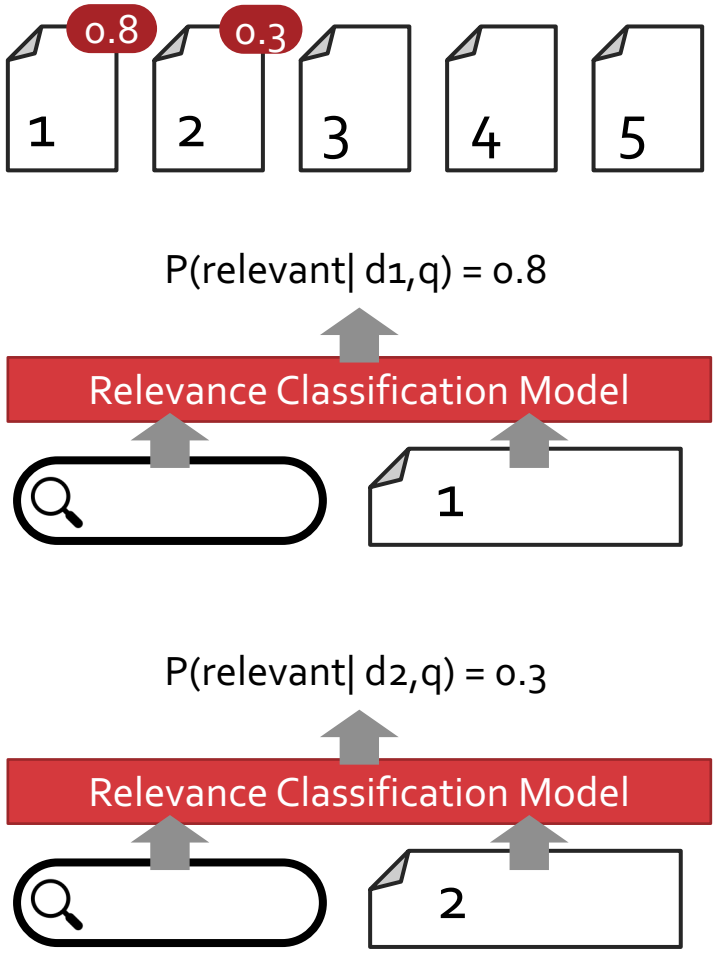


Current Retrieval Systems

Dense Retrieval

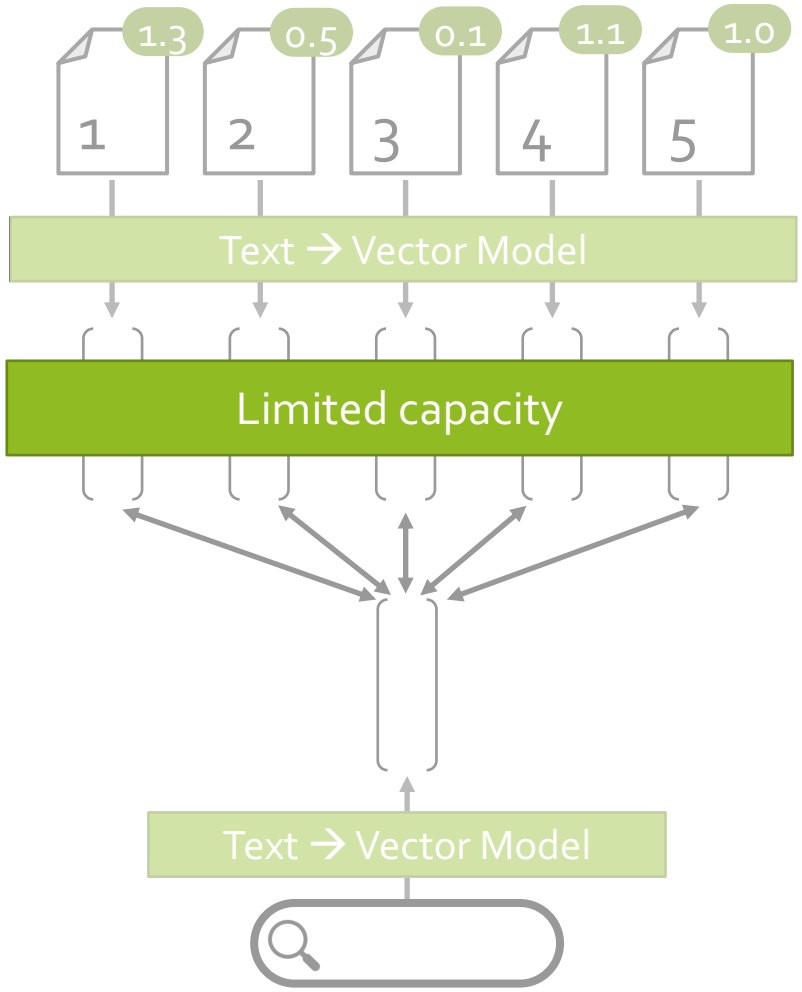


Cross-Encoders

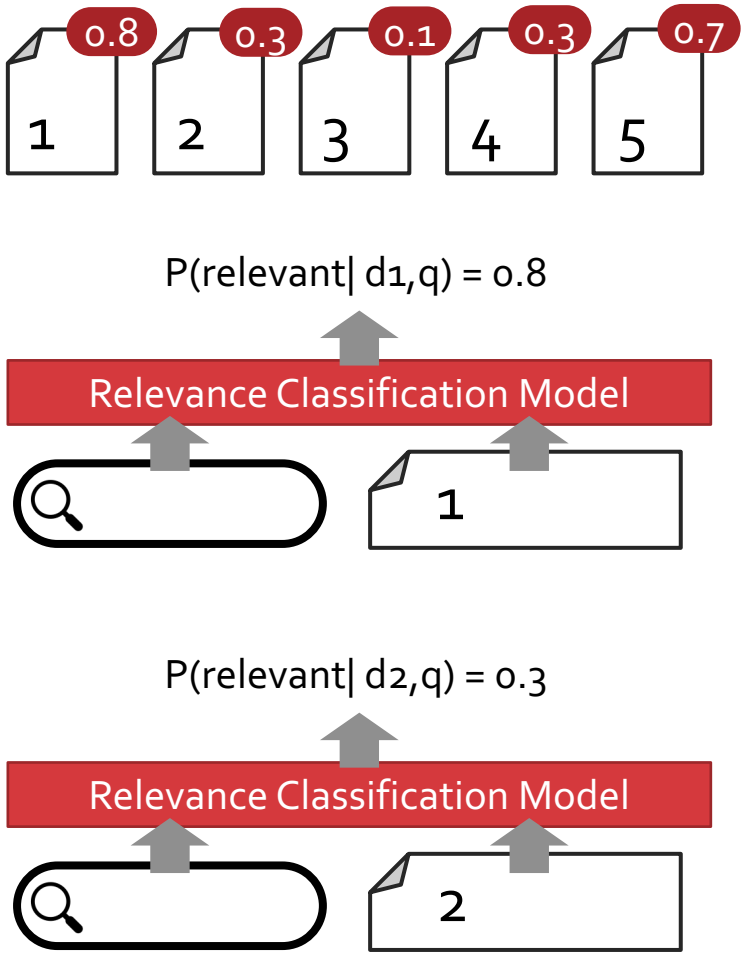


Current Retrieval Systems

Dense Retrieval

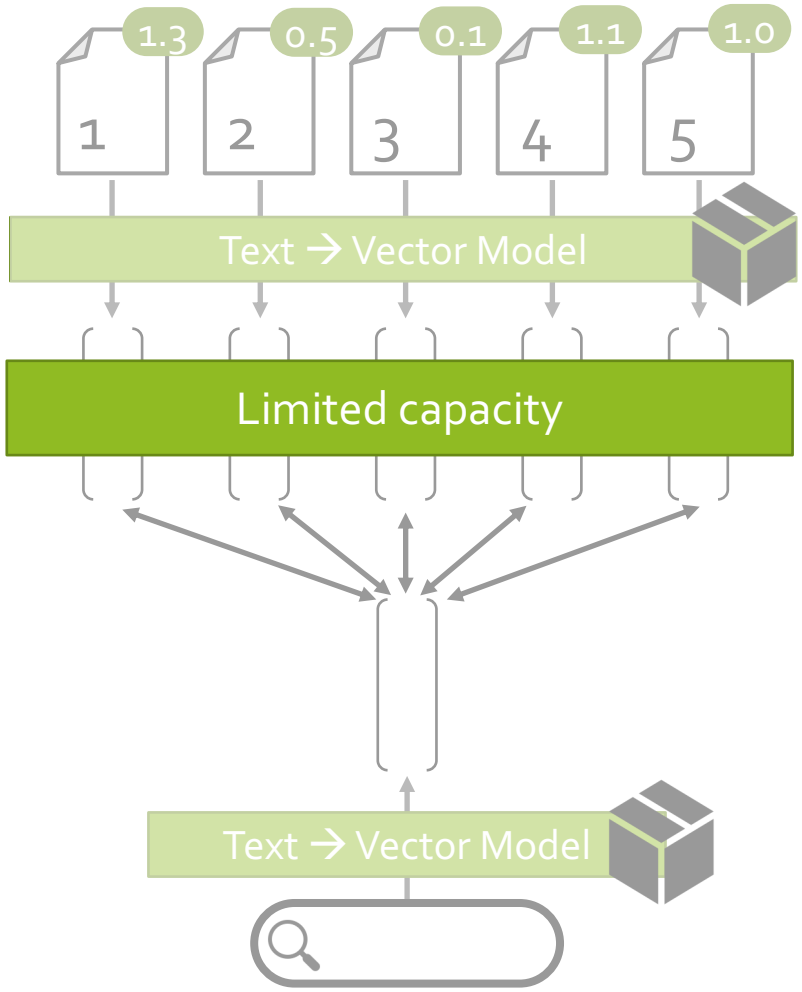


Cross-Encoders

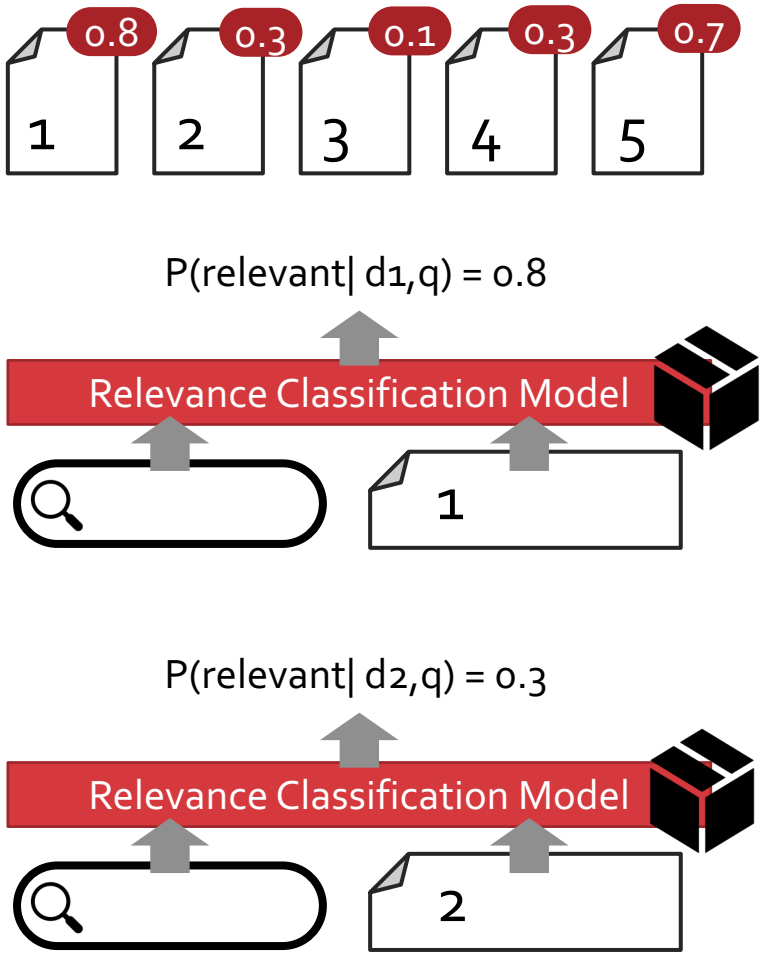


Current Retrieval Systems

Dense Retrieval

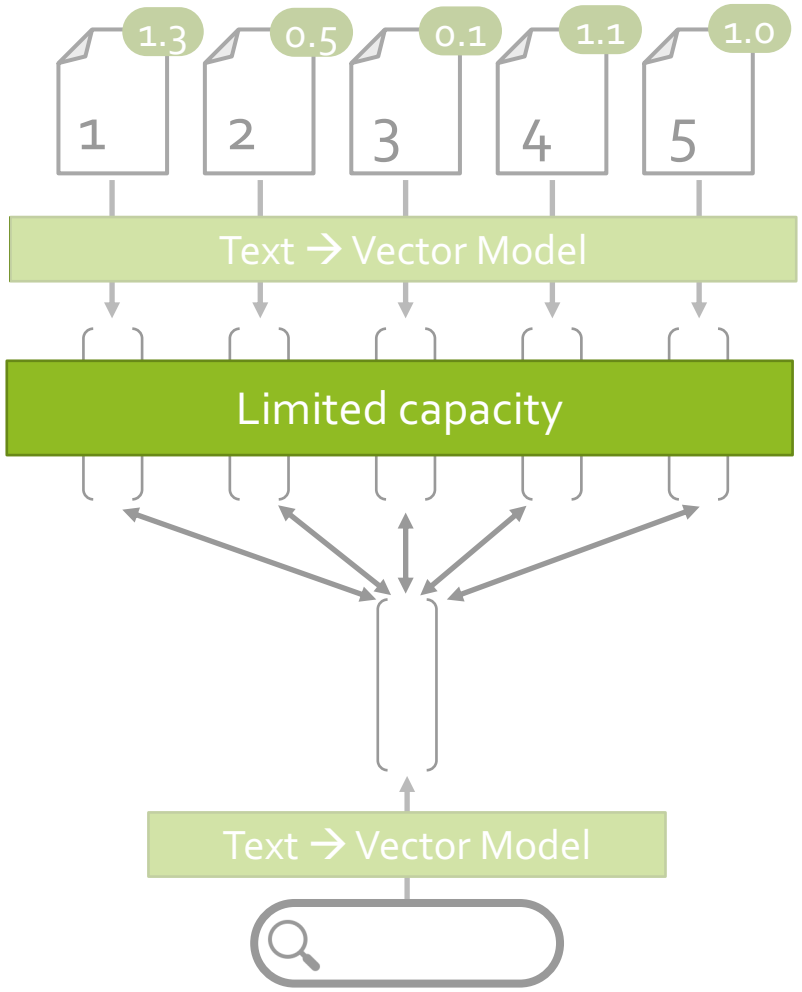


Cross-Encoders

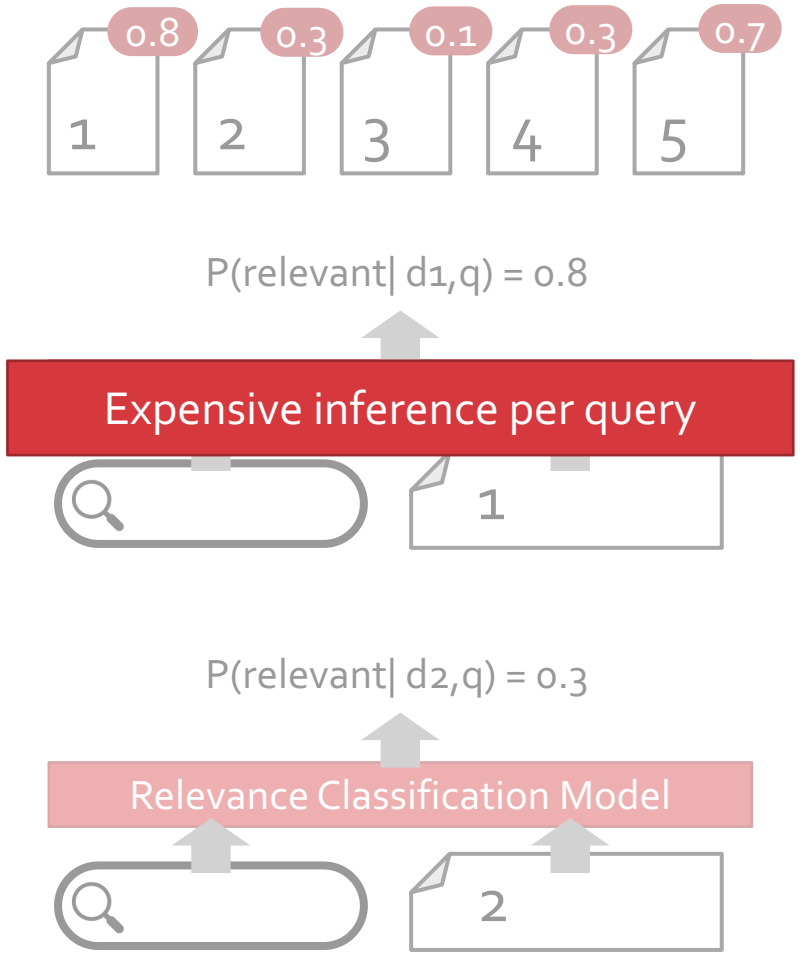


Current Retrieval Systems

Dense Retrieval

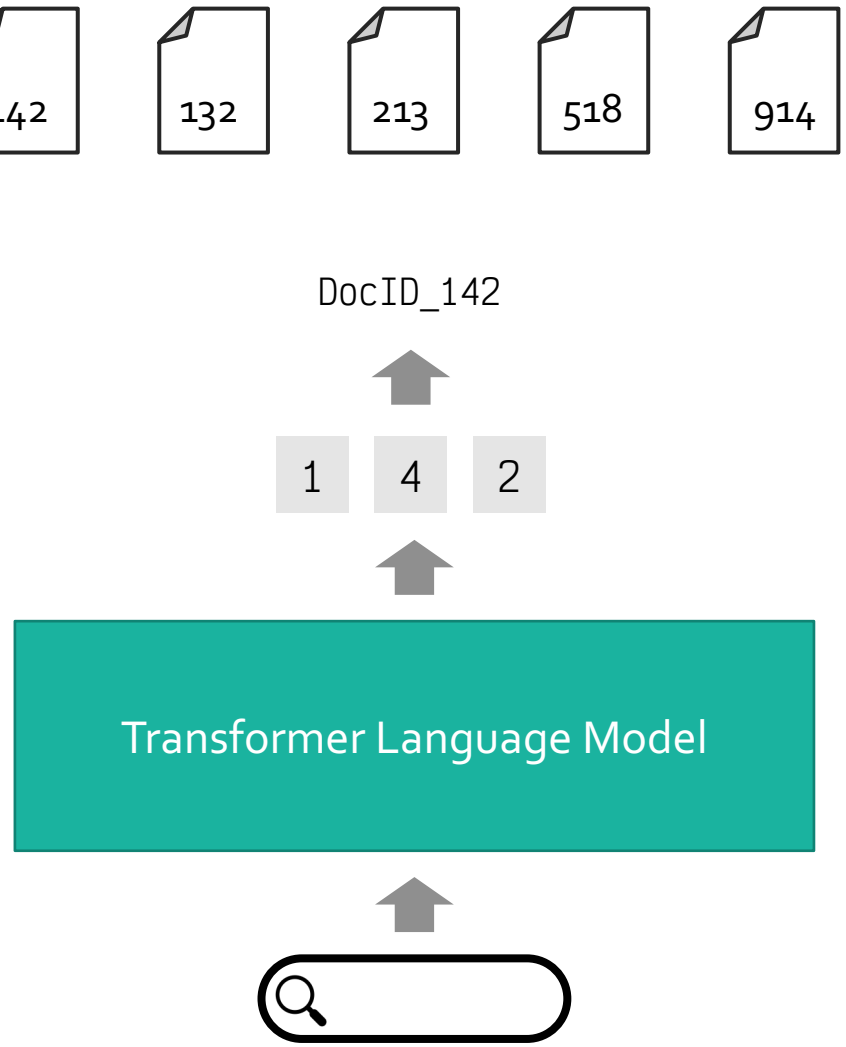


Cross-Encoders



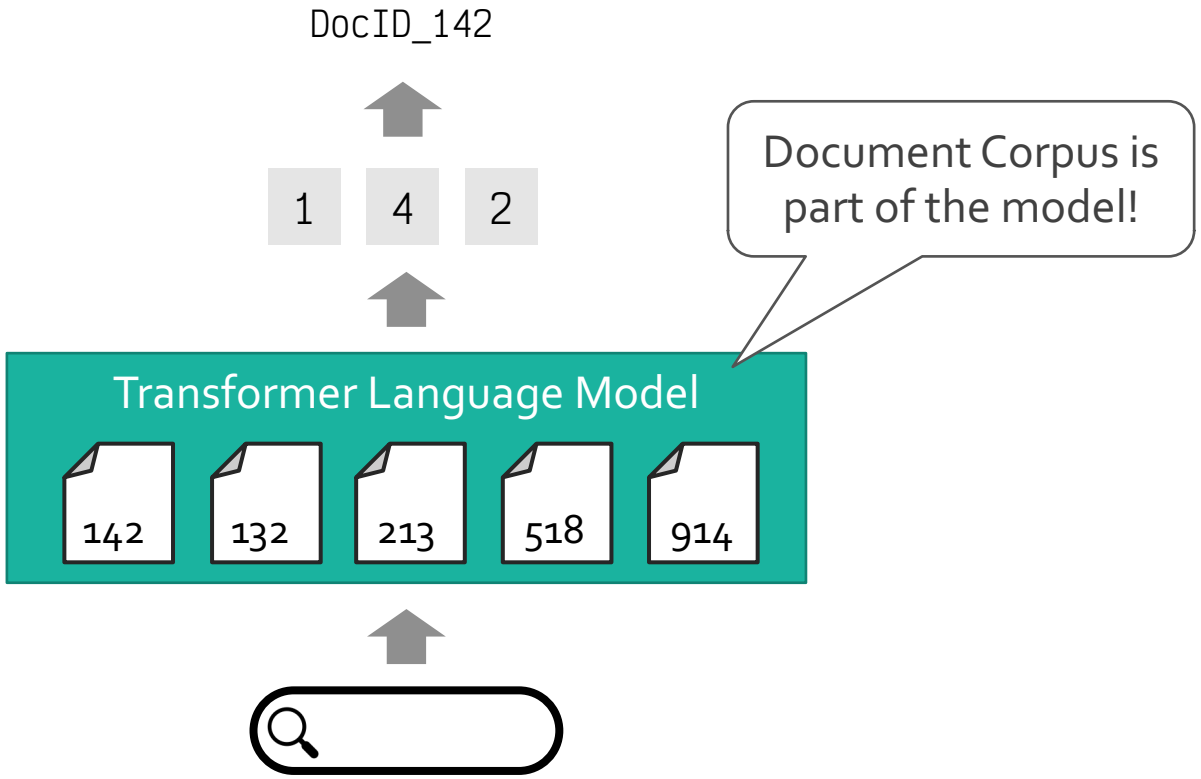
Generative Information Retrieval

Generative Information Retrieval



Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

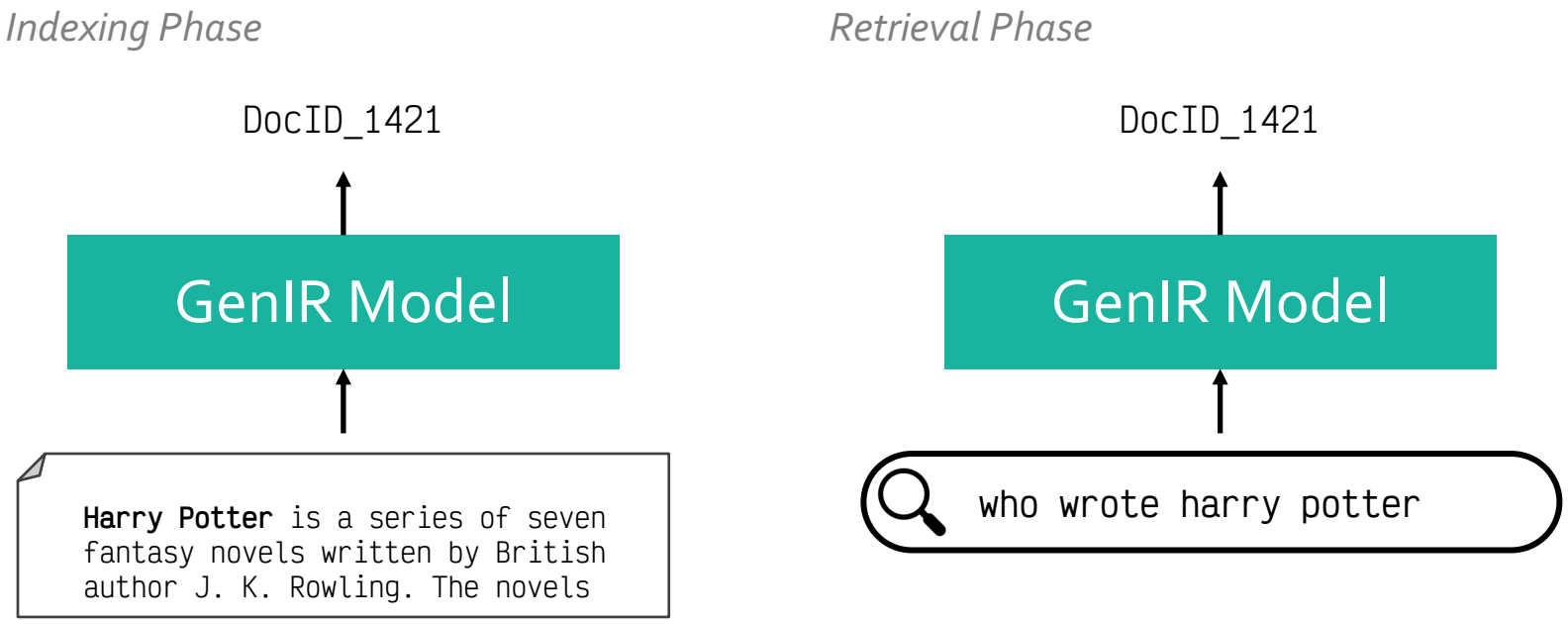
Generative Information Retrieval



Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

GenIR Training

Goal: Associate the content of a document with its ID



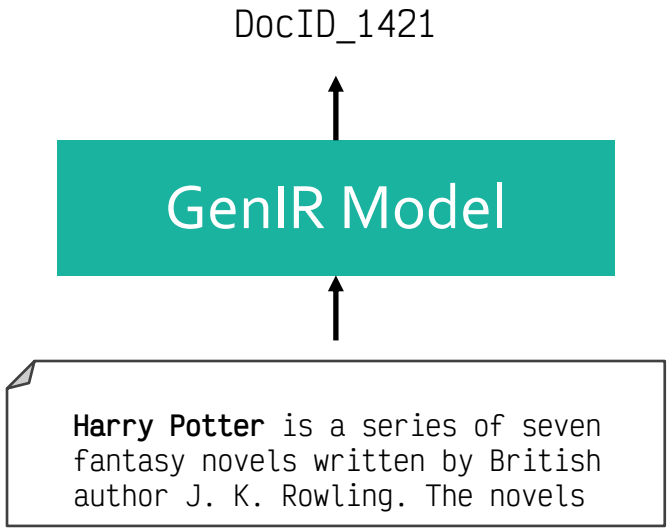
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

GenIR Training

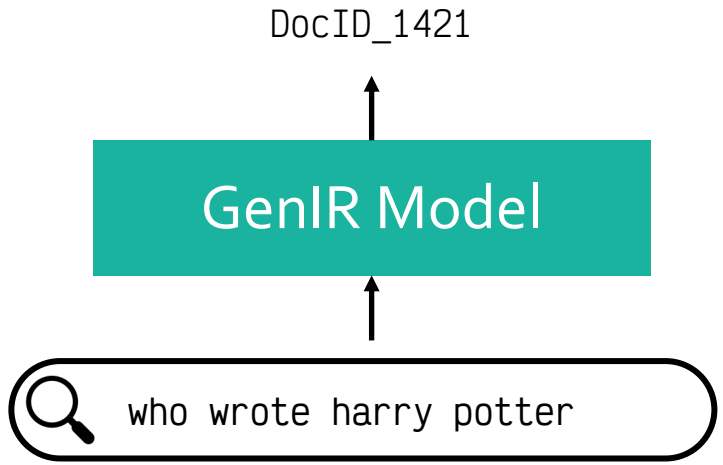
Goal: Associate the content of a document with its ID

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

Indexing Phase



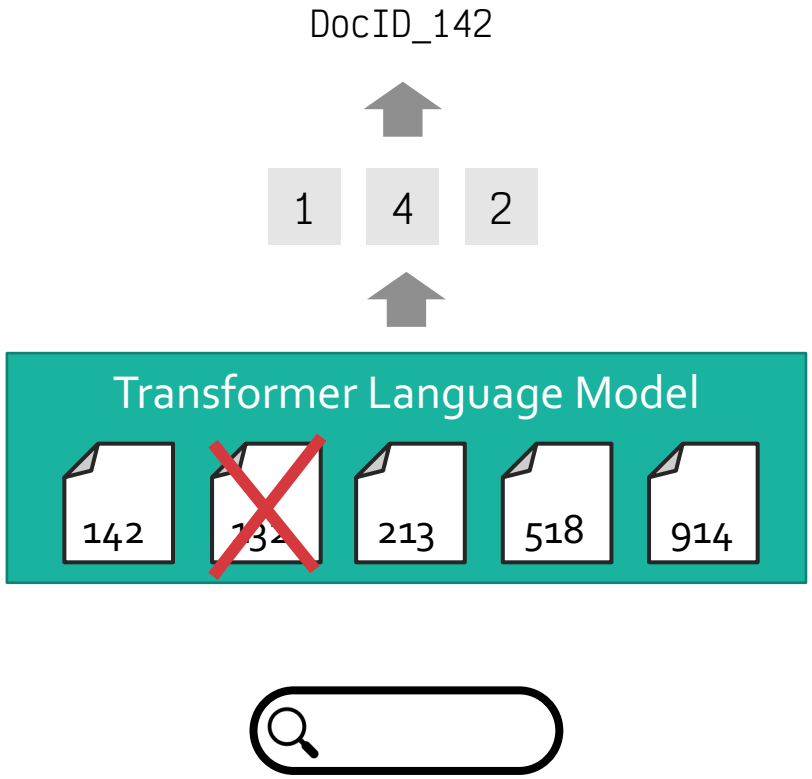
Retrieval Phase



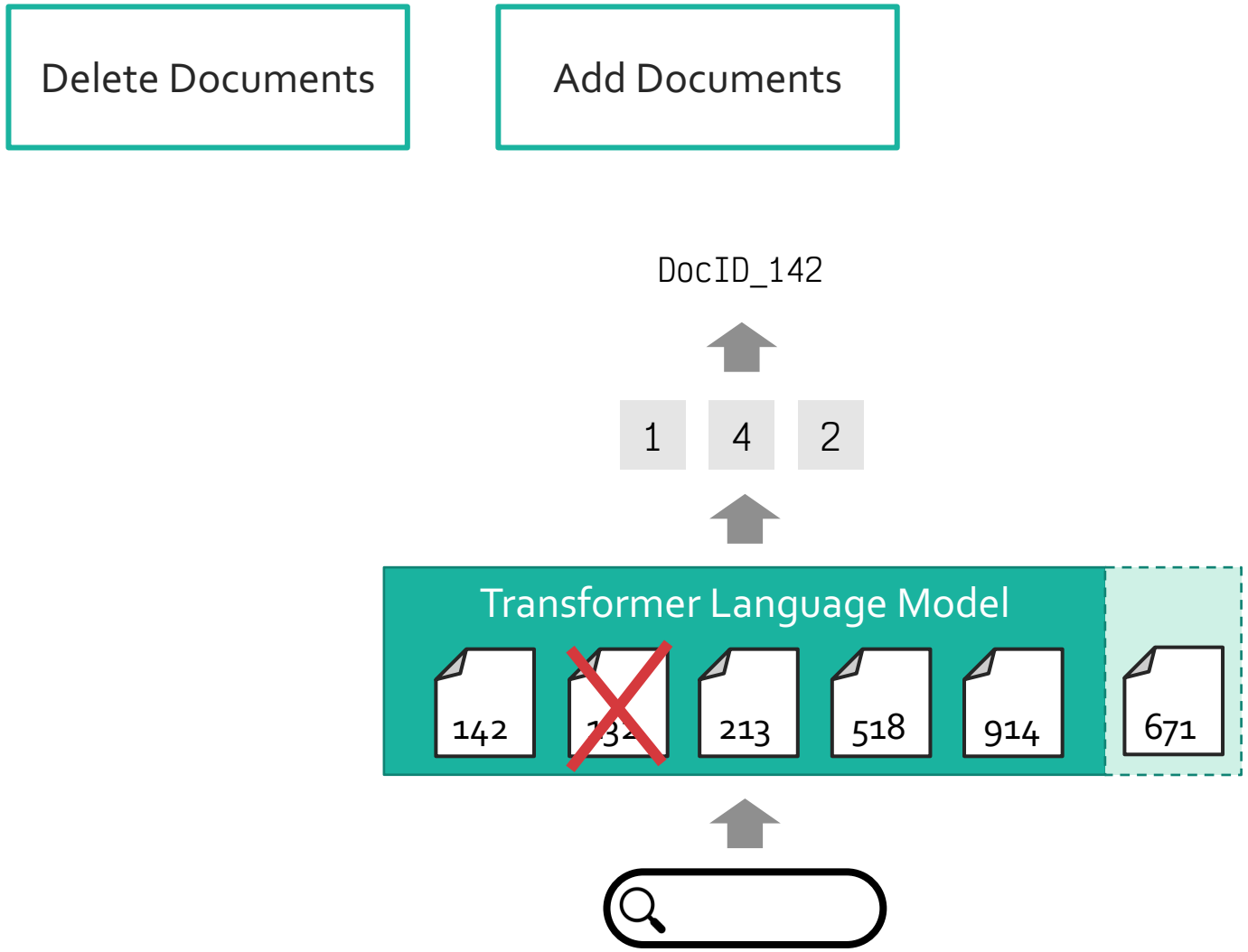
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

The Problem

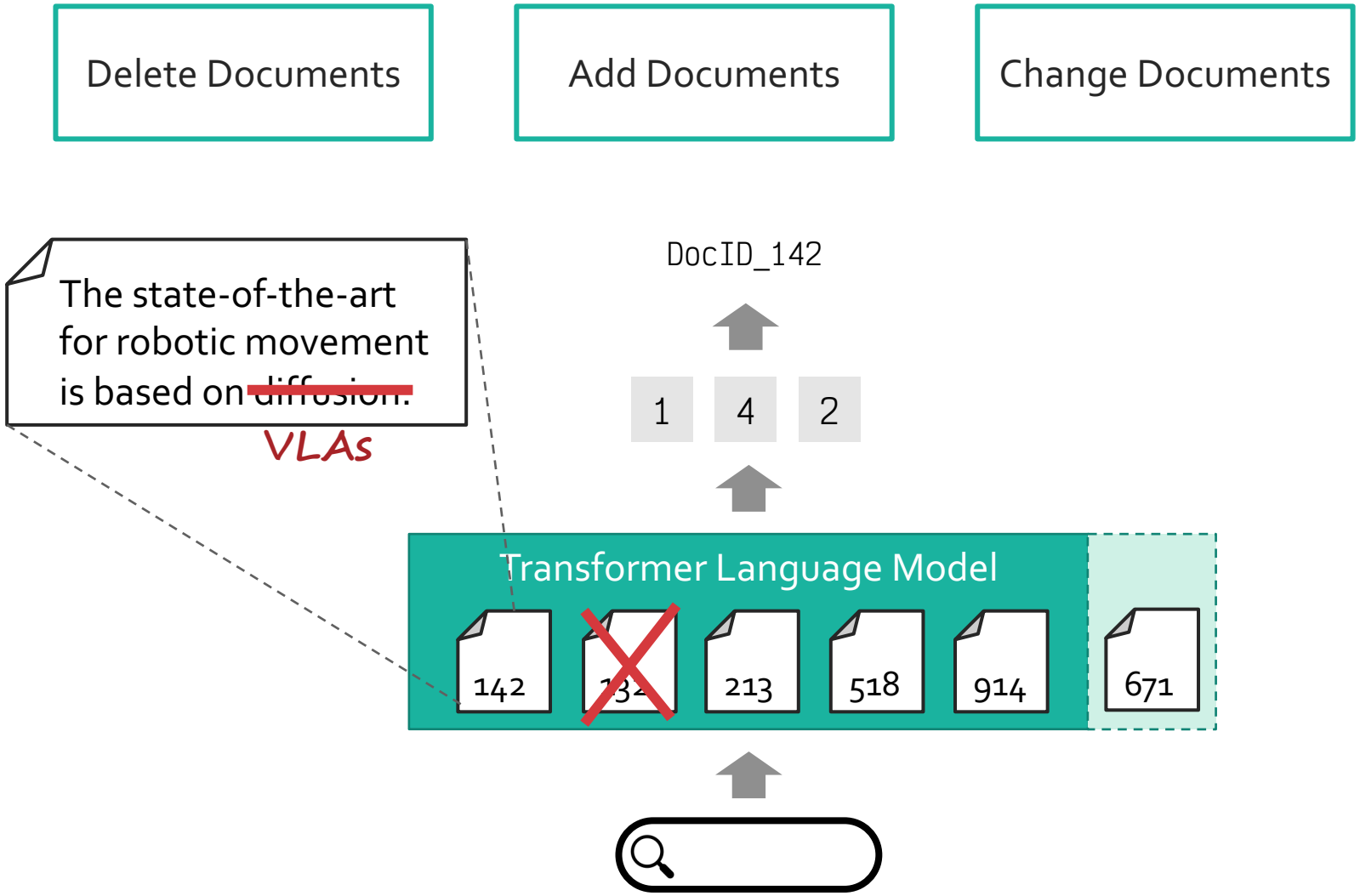
Delete Documents



The Problem


Delete Documents
Add Documents
DocID_142
1
4
2
Transformer Language Model
142
213
518
914
671

The Problem


Delete Documents
Add Documents
Change Documents
The state-of-the-art for robotic movement is based on diffusion.
VLAs
DocID_142
1
4
2
Transformer Language Model
142
213
518
914
671

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

TECHNION

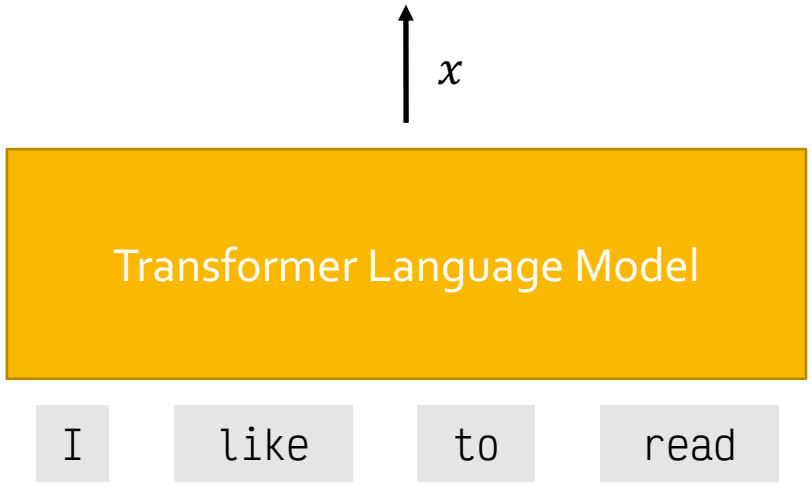
Funding:

Funded by the European Union

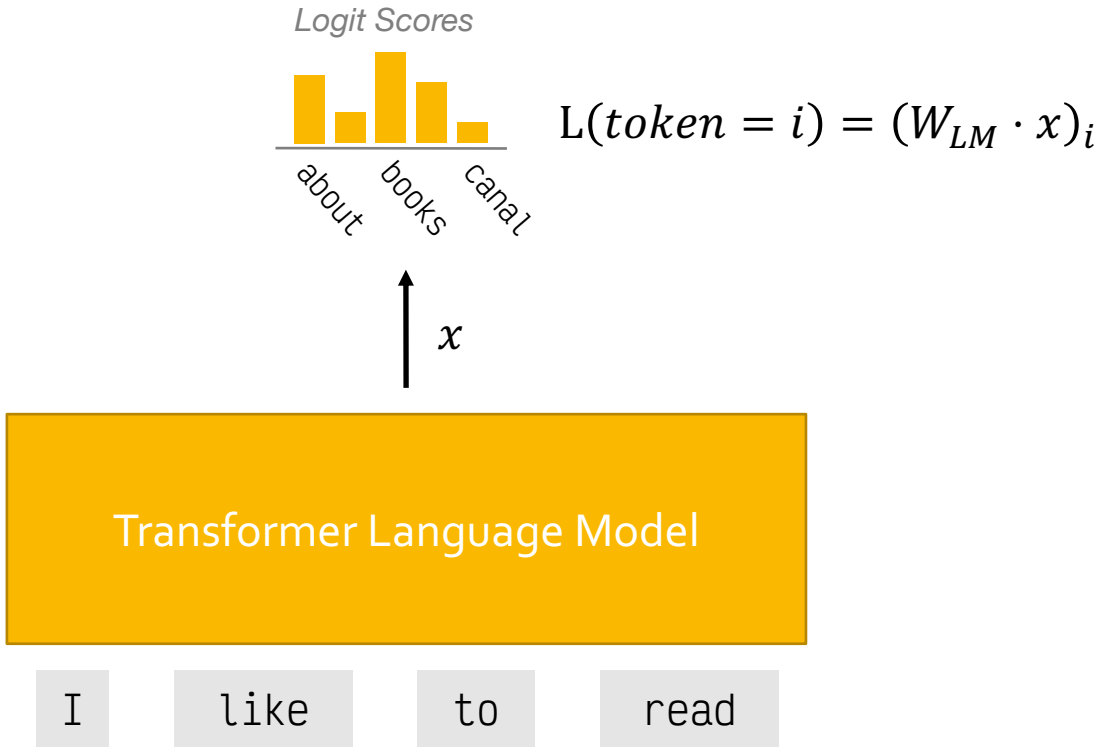
European Research Council
Established by the European Commission

Fondation Azrieli Foundation

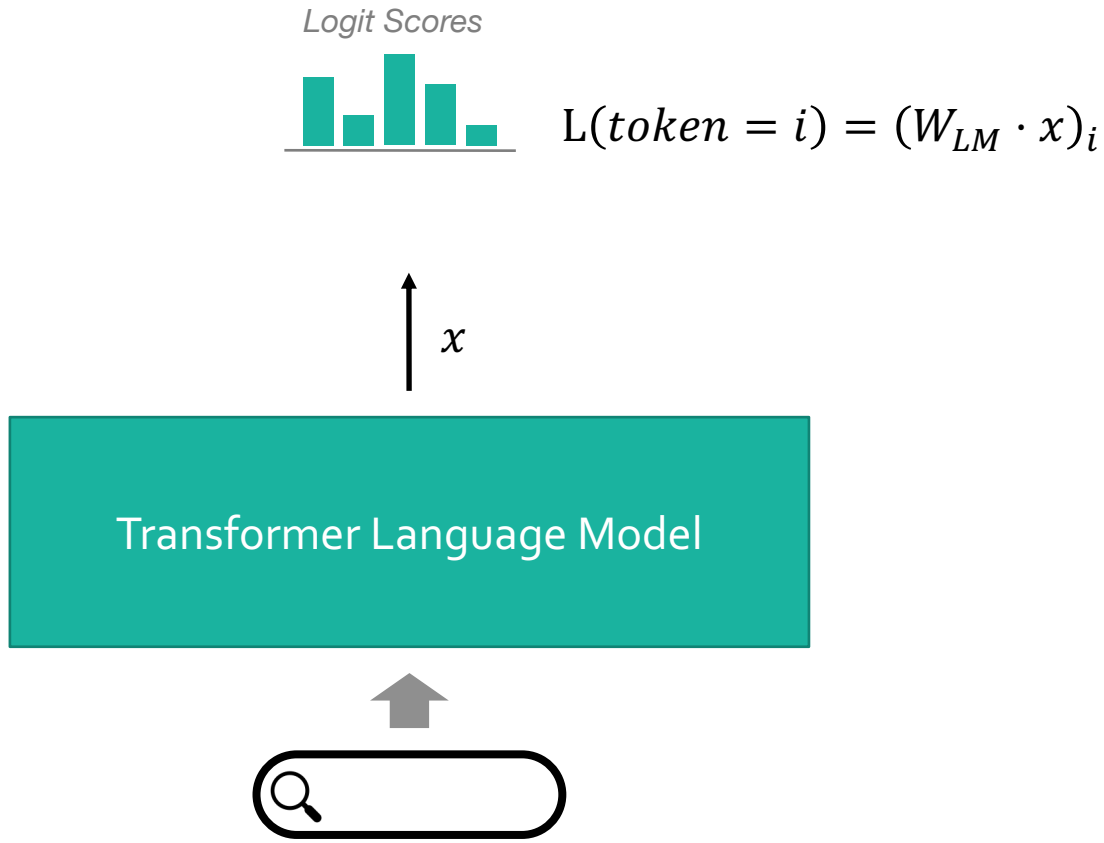
Generative Models



Generative Models

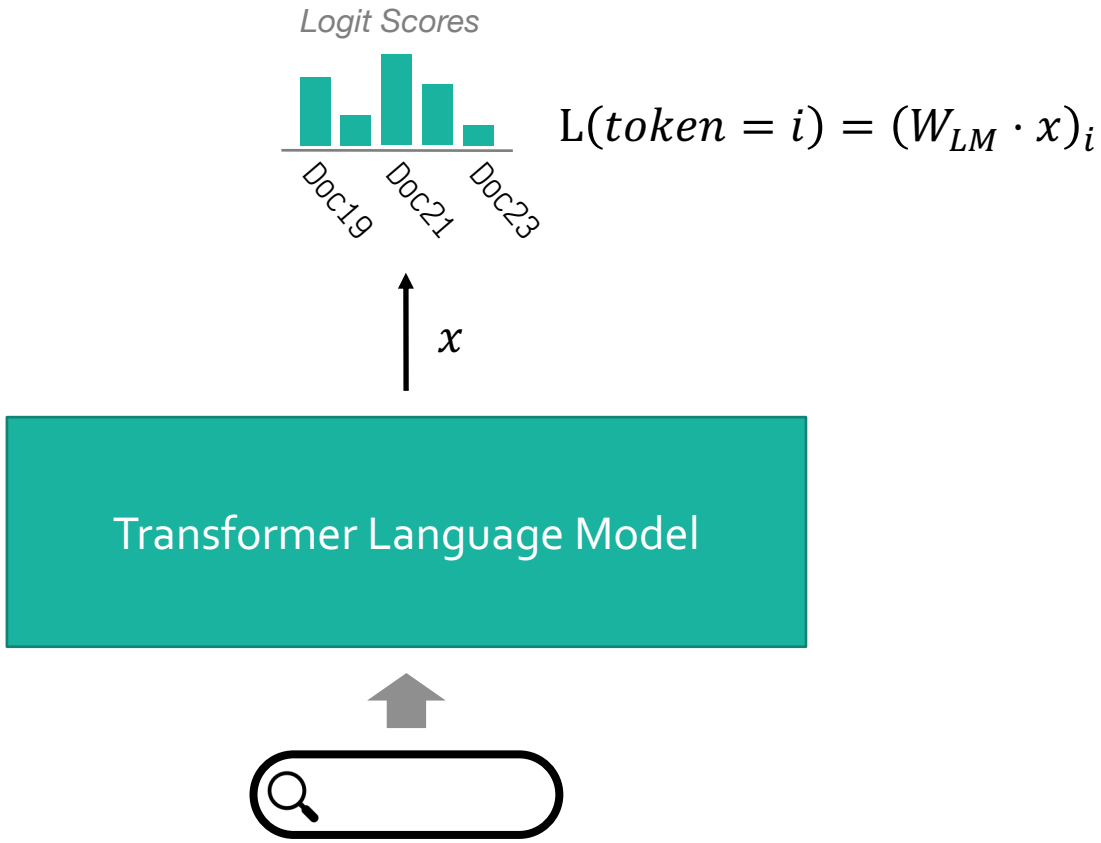


Generative Information Retrieval



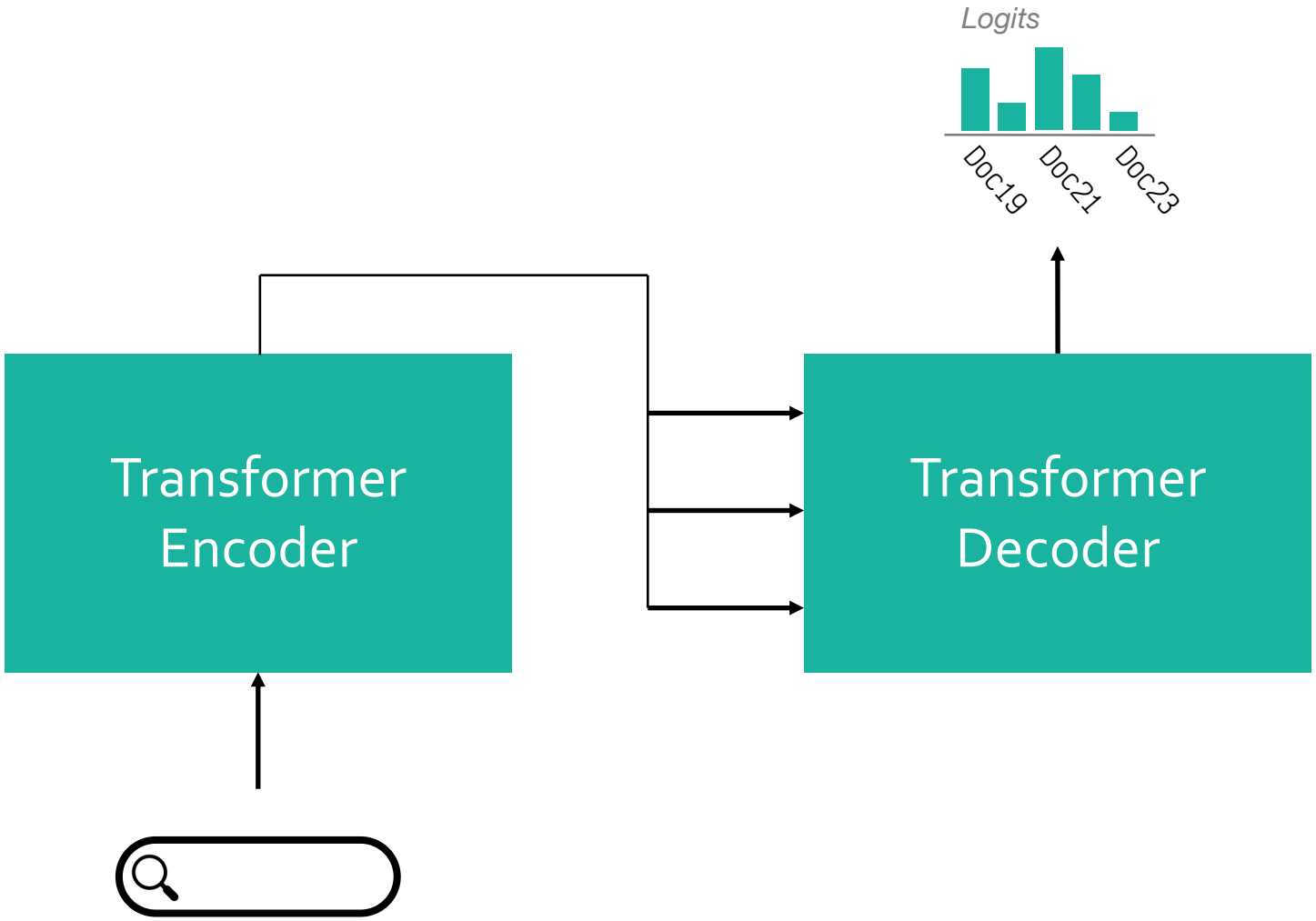
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

Generative Information Retrieval



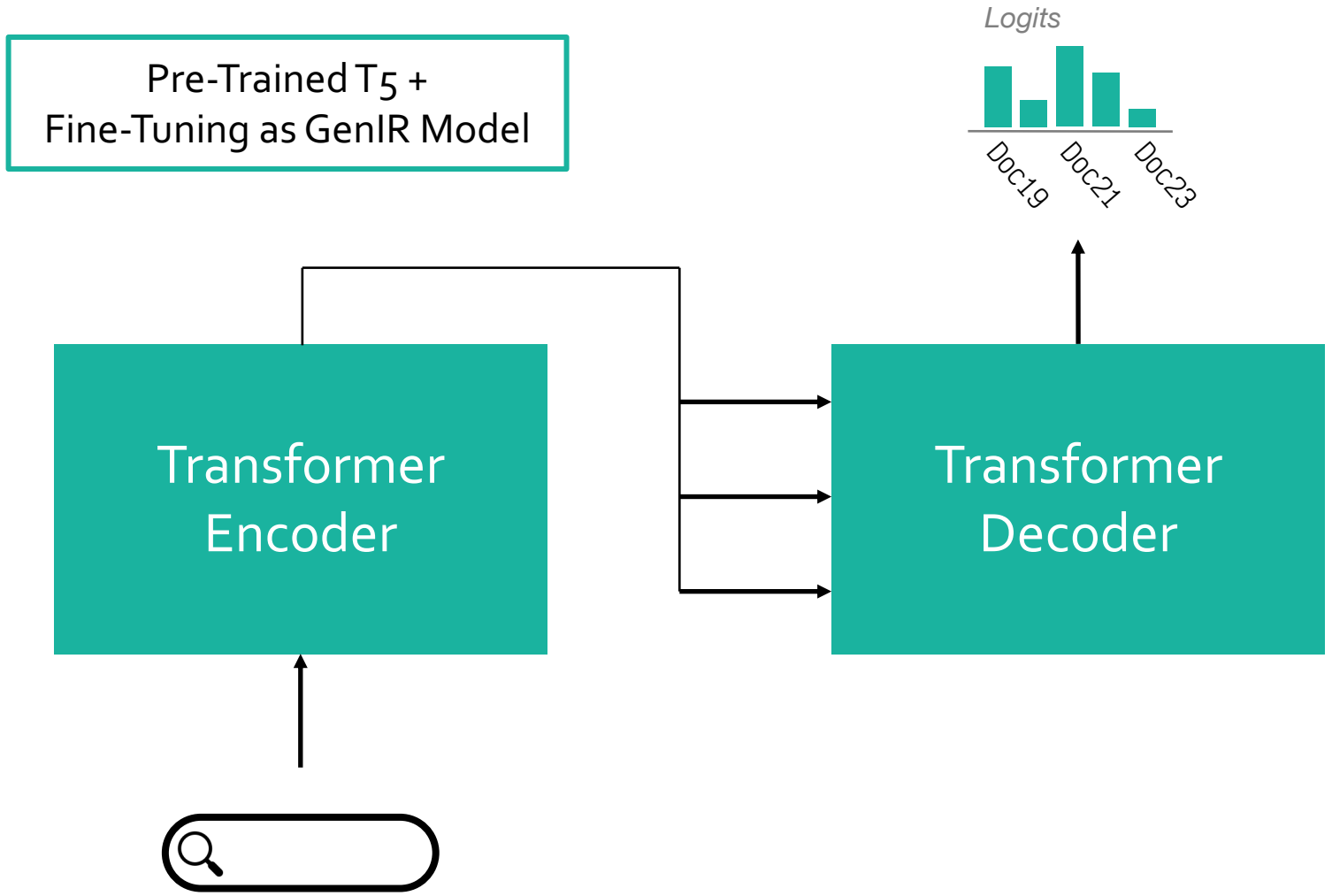
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

Generative Encoder-Decoder



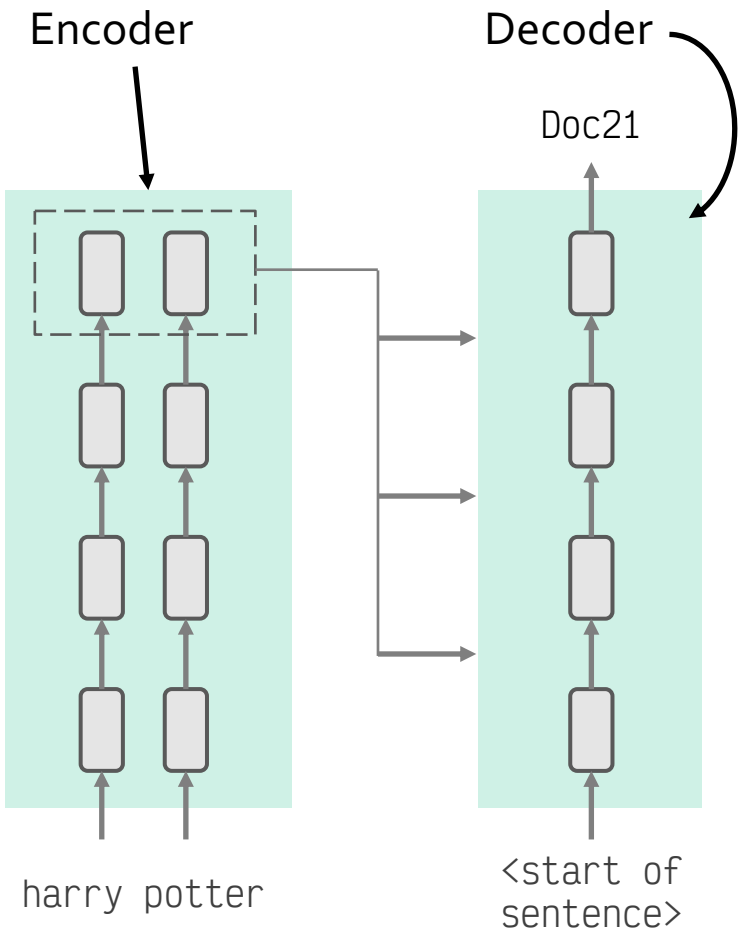
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

Generative Encoder-Decoder



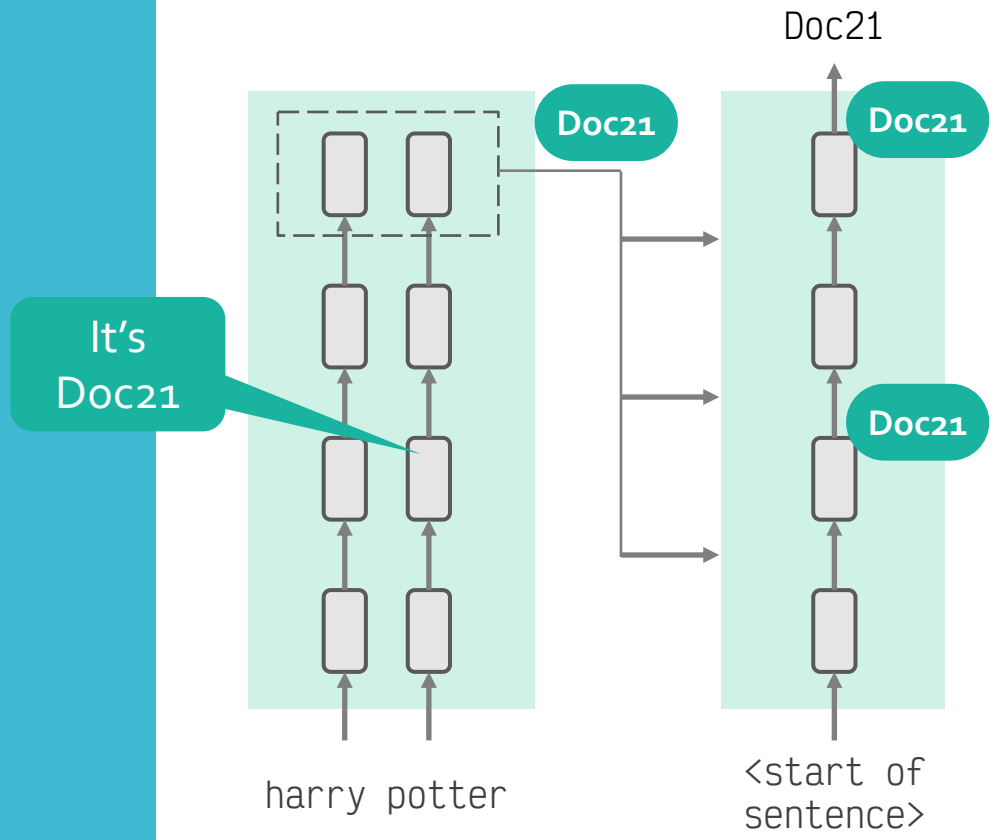
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

Which part identifies the document?

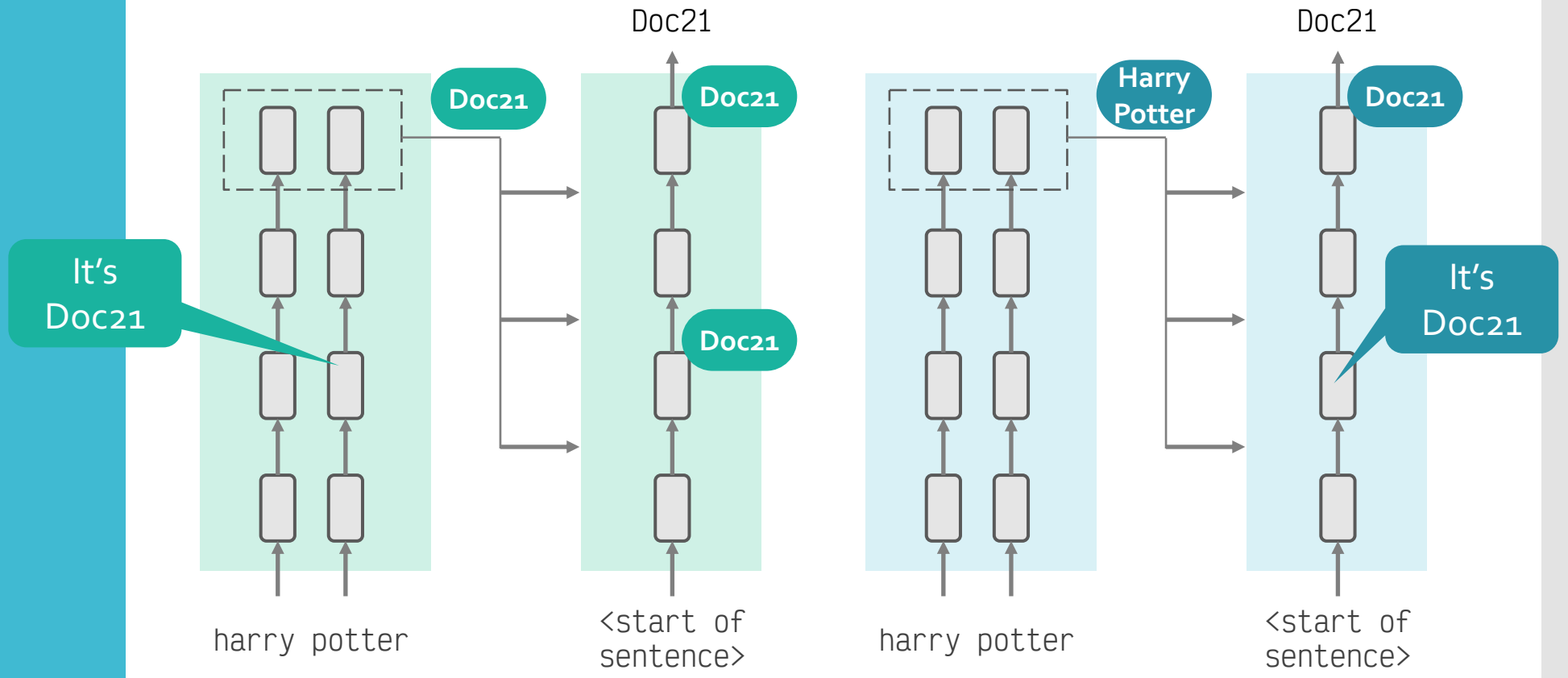


Two Hypotheses

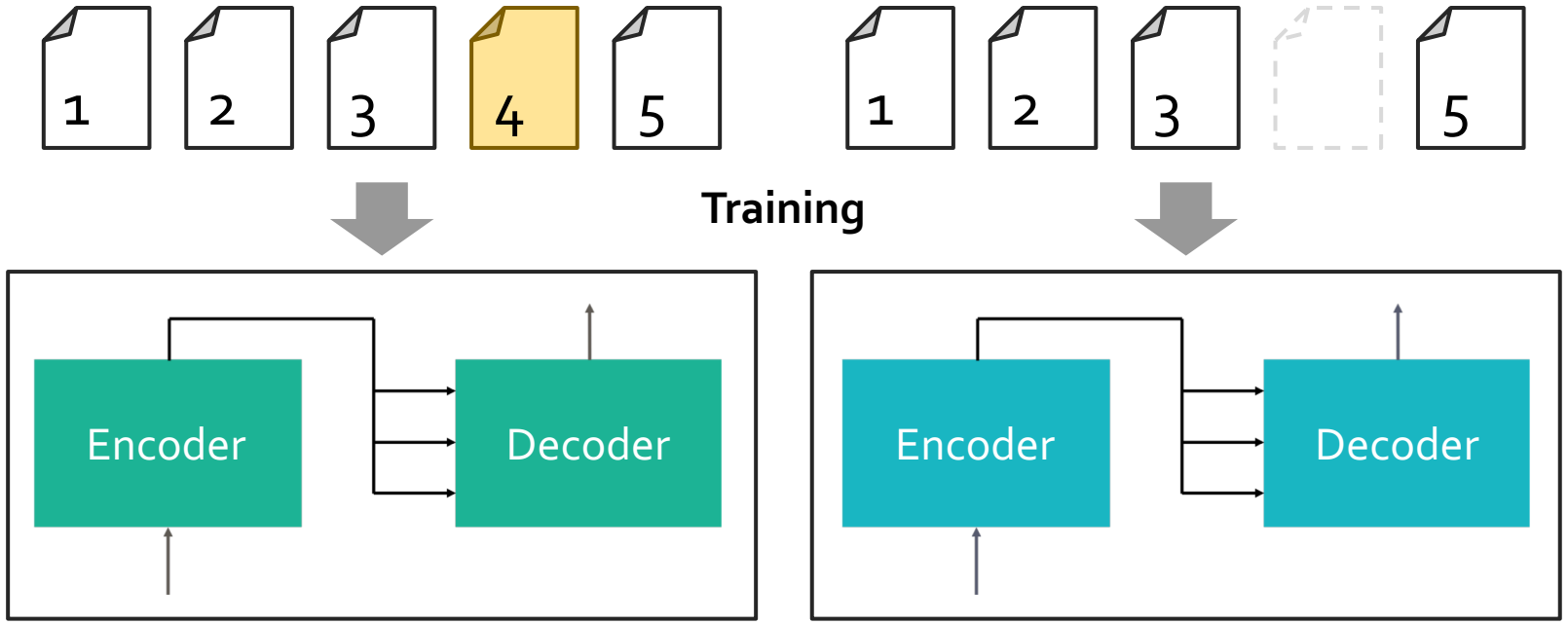
Encoder identifies the document



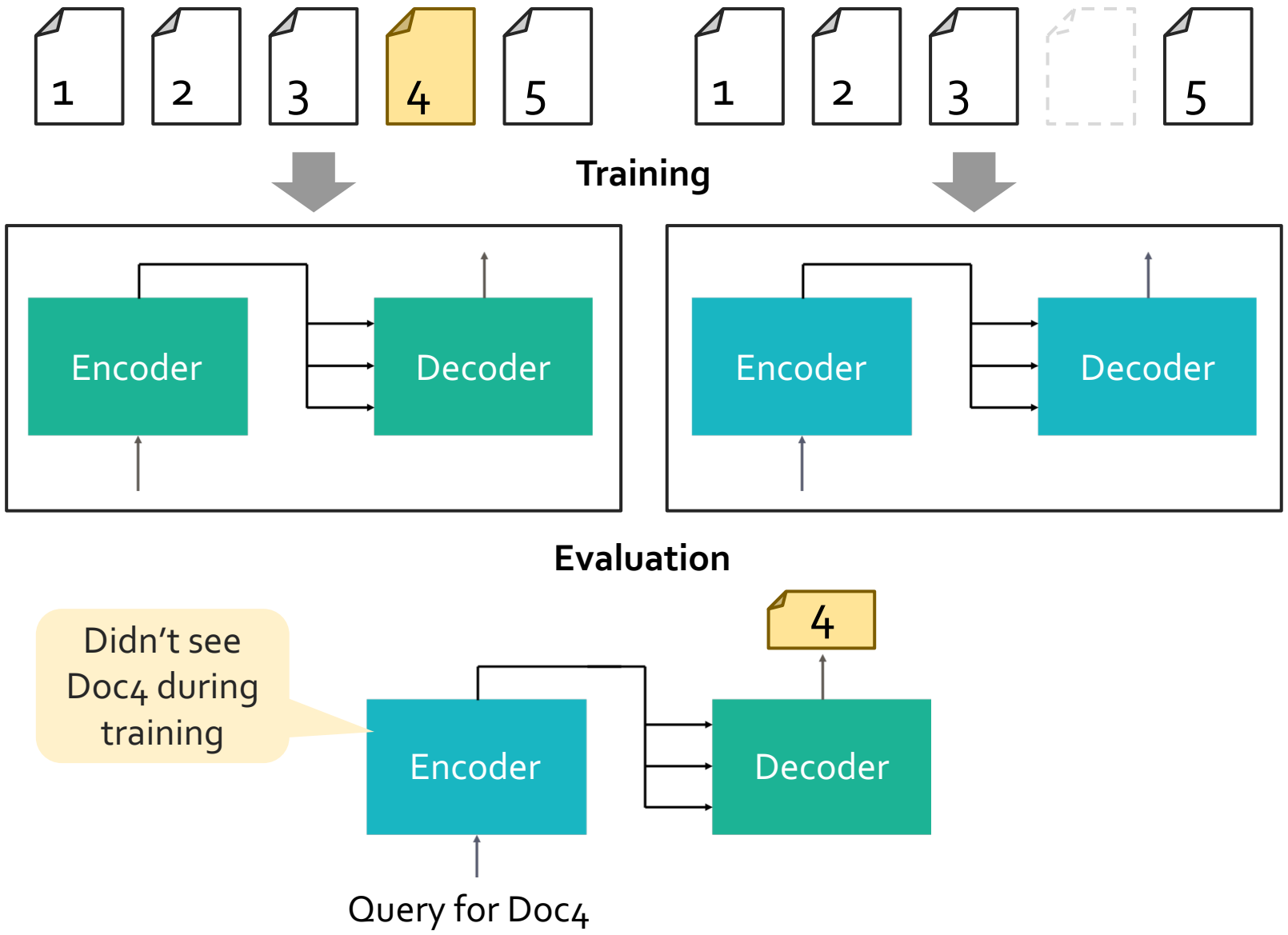
Two Hypotheses



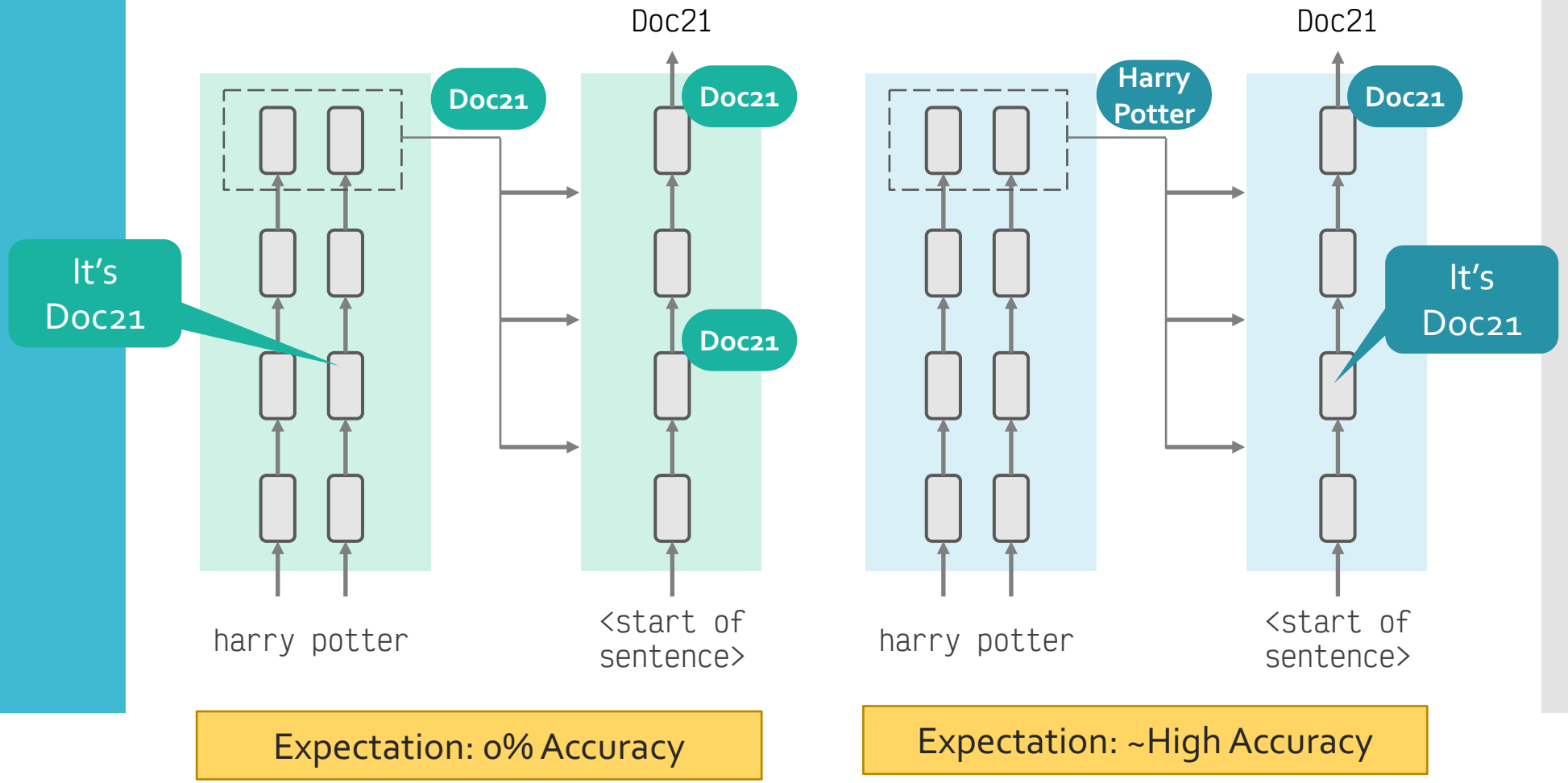
Looking at the Encoder



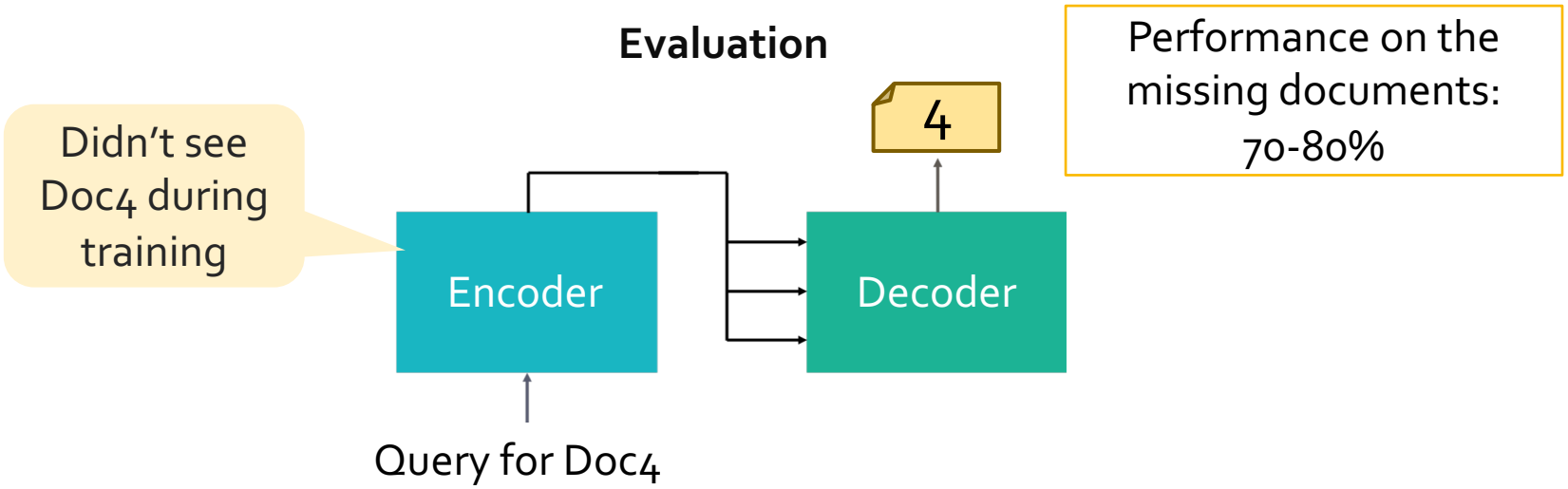
Looking at the Encoder



Two Hypotheses



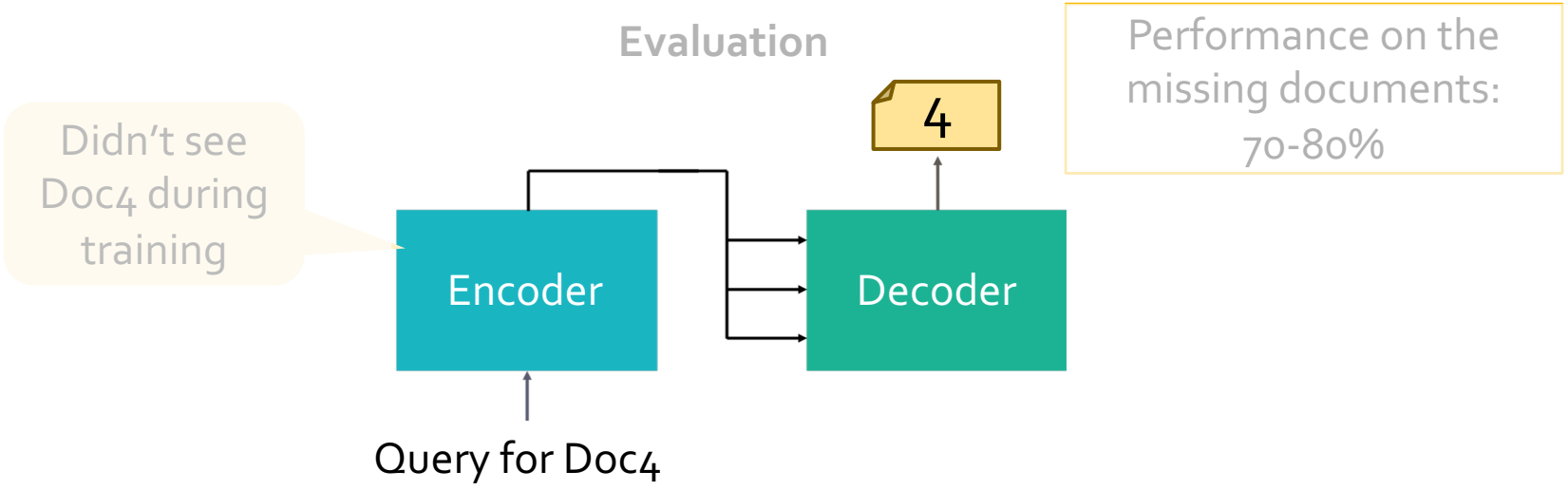
Looking at the Encoder



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%

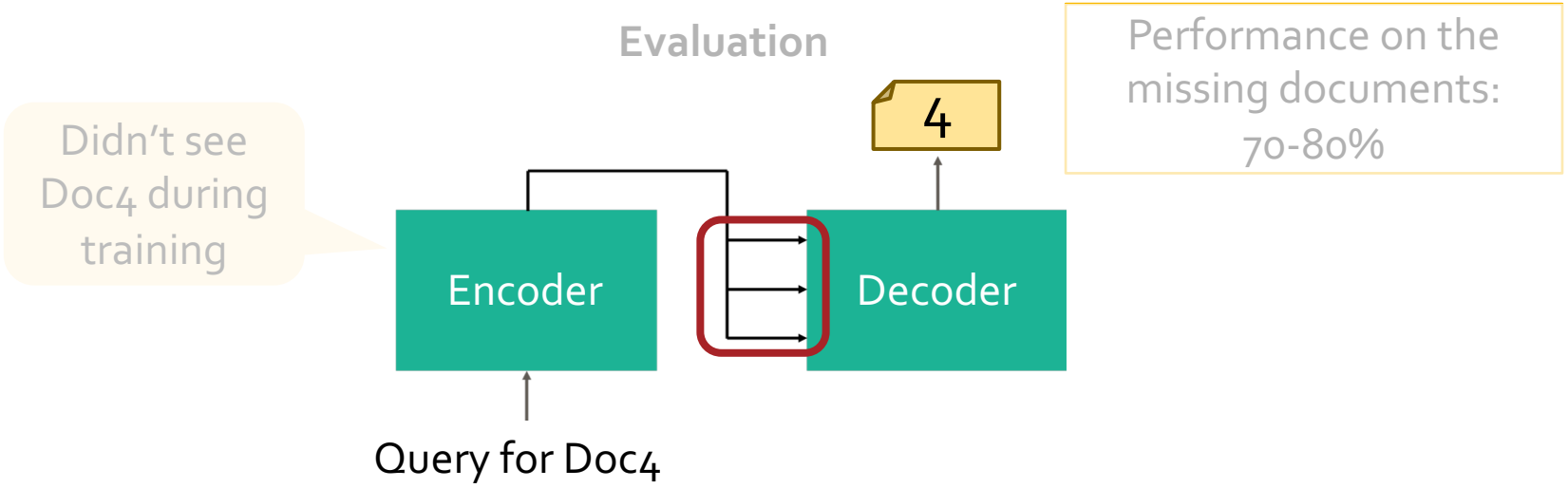
Looking at the Encoder



Conclusion: The decoder identifies the document.

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

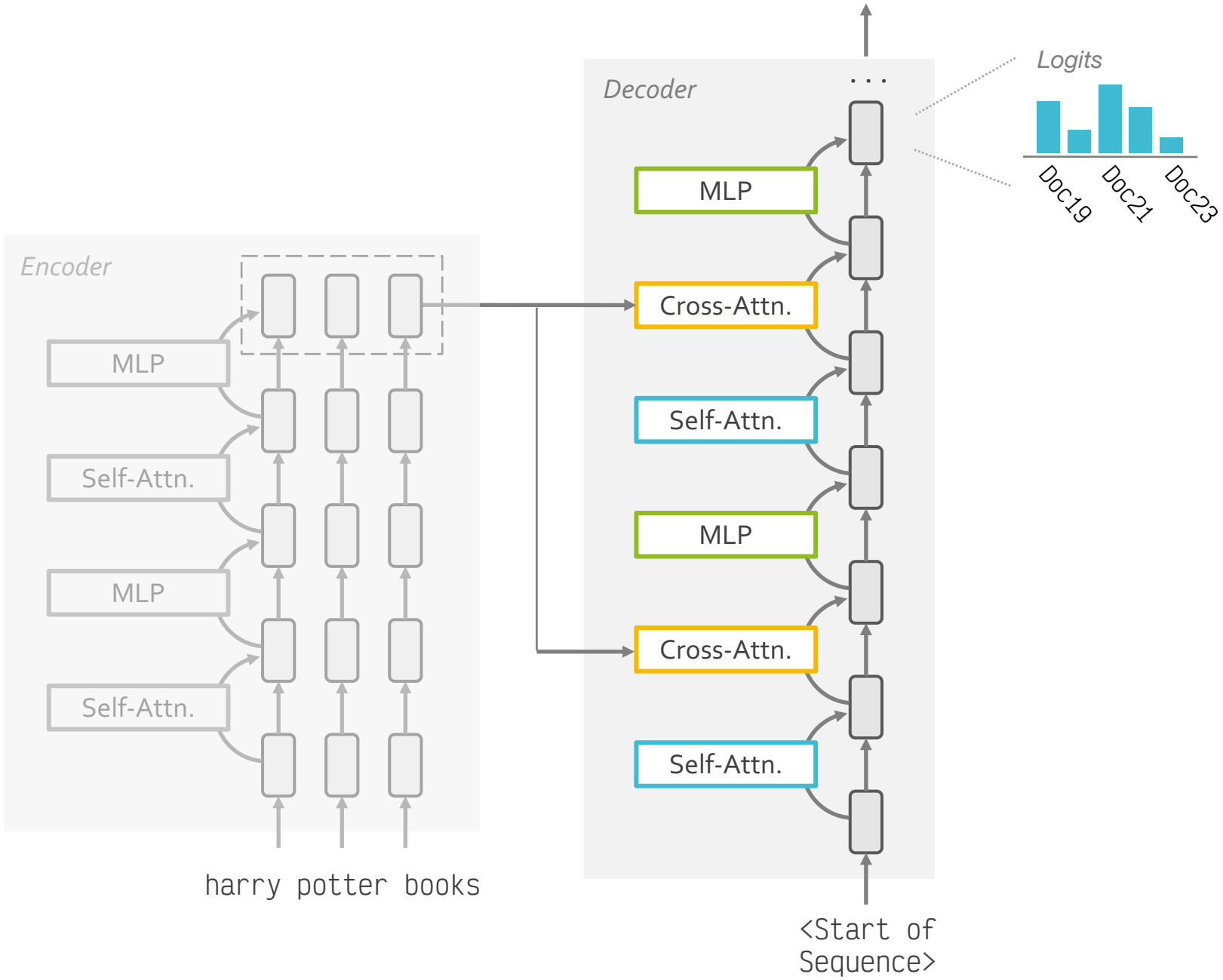
Looking at the Encoder



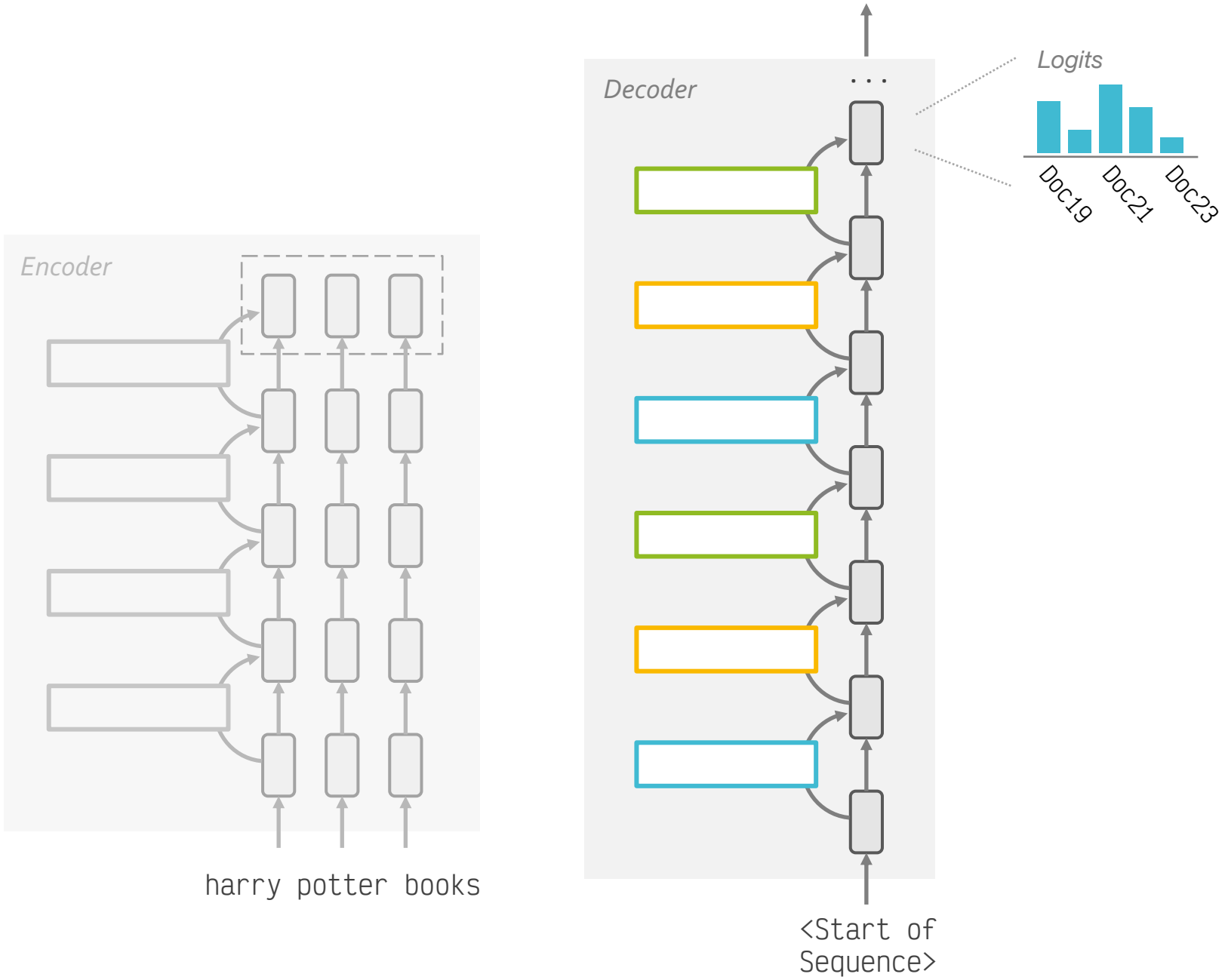
Setup

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

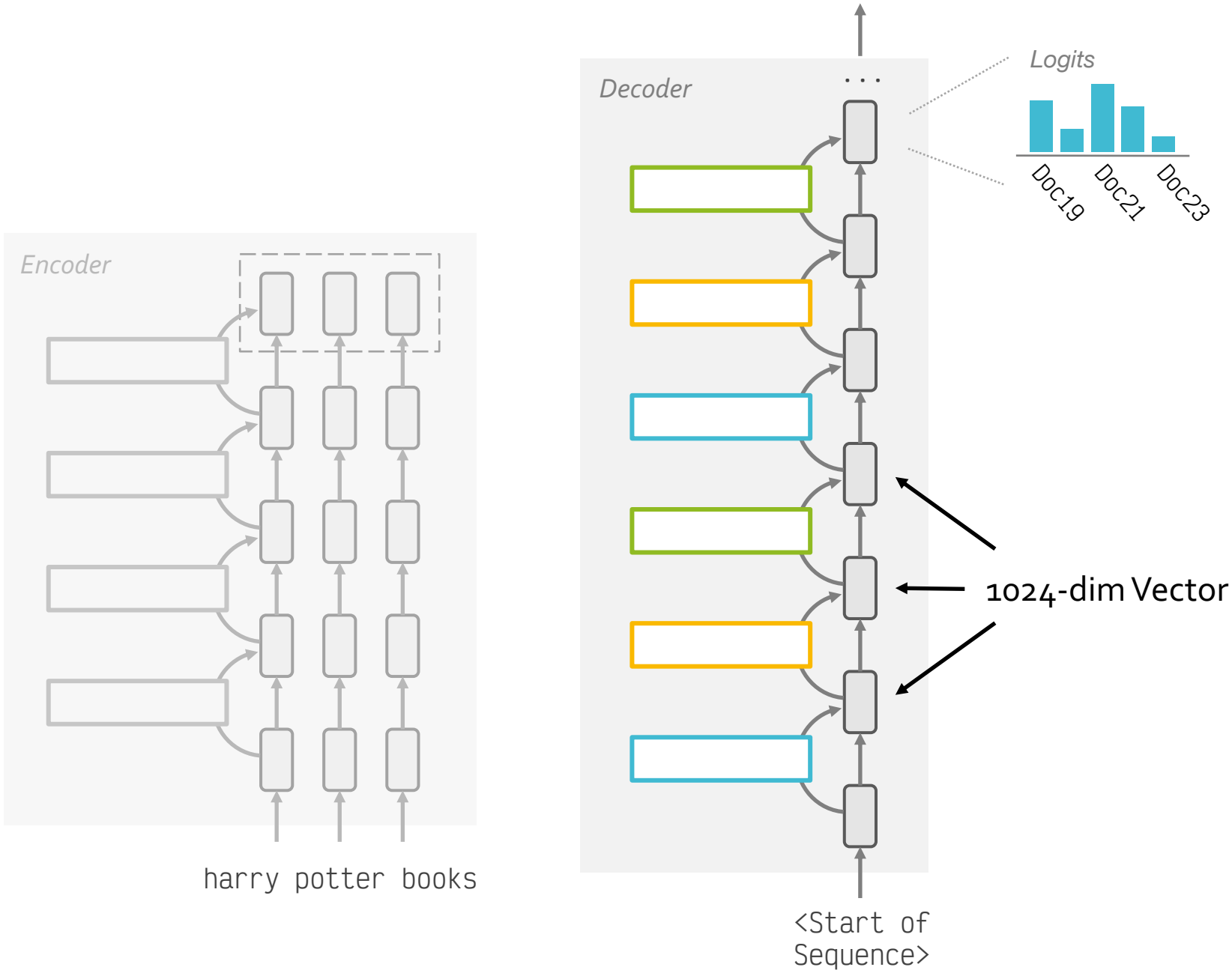
How does a GenIR Model look like?



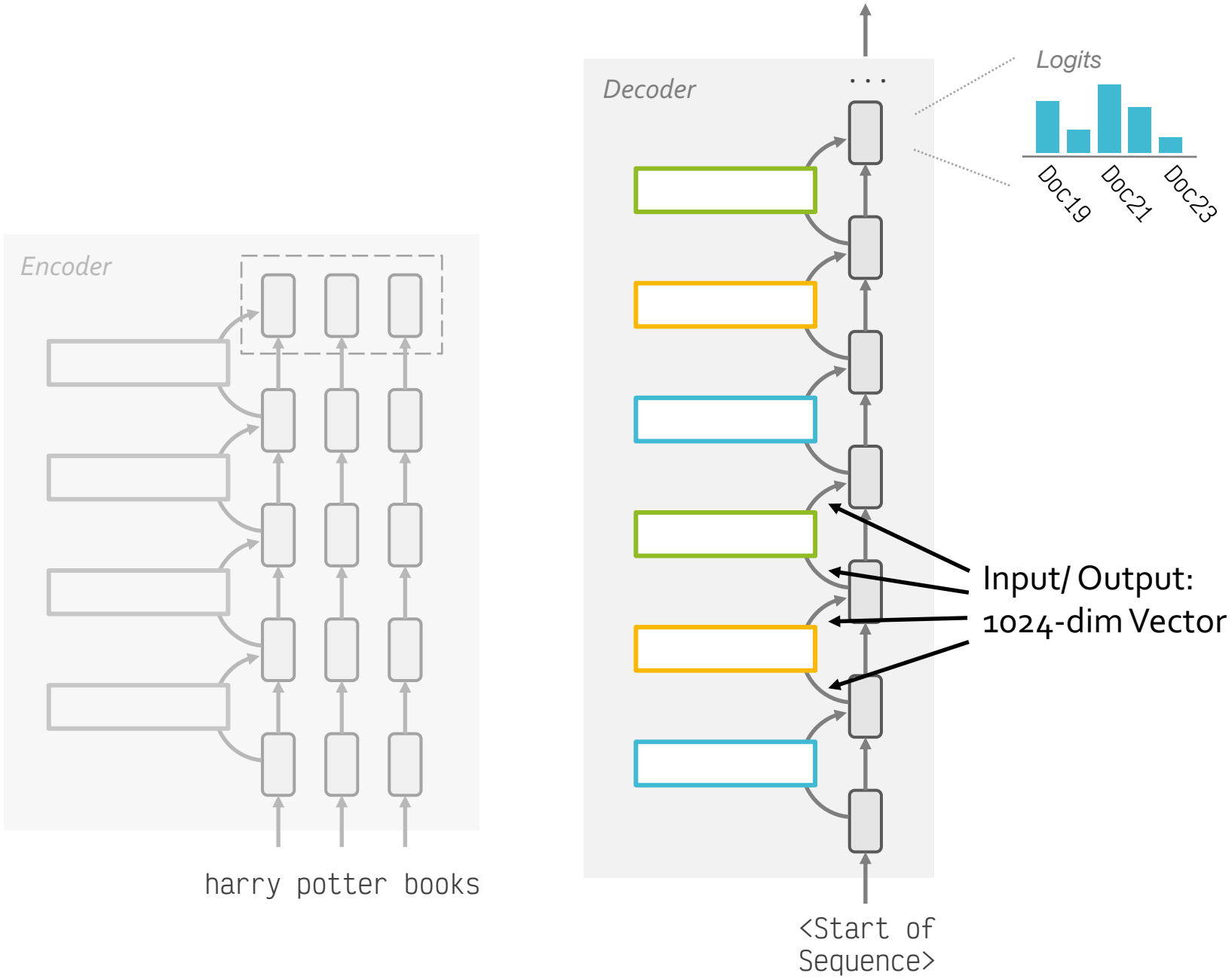
How does a GenIR Model look like?



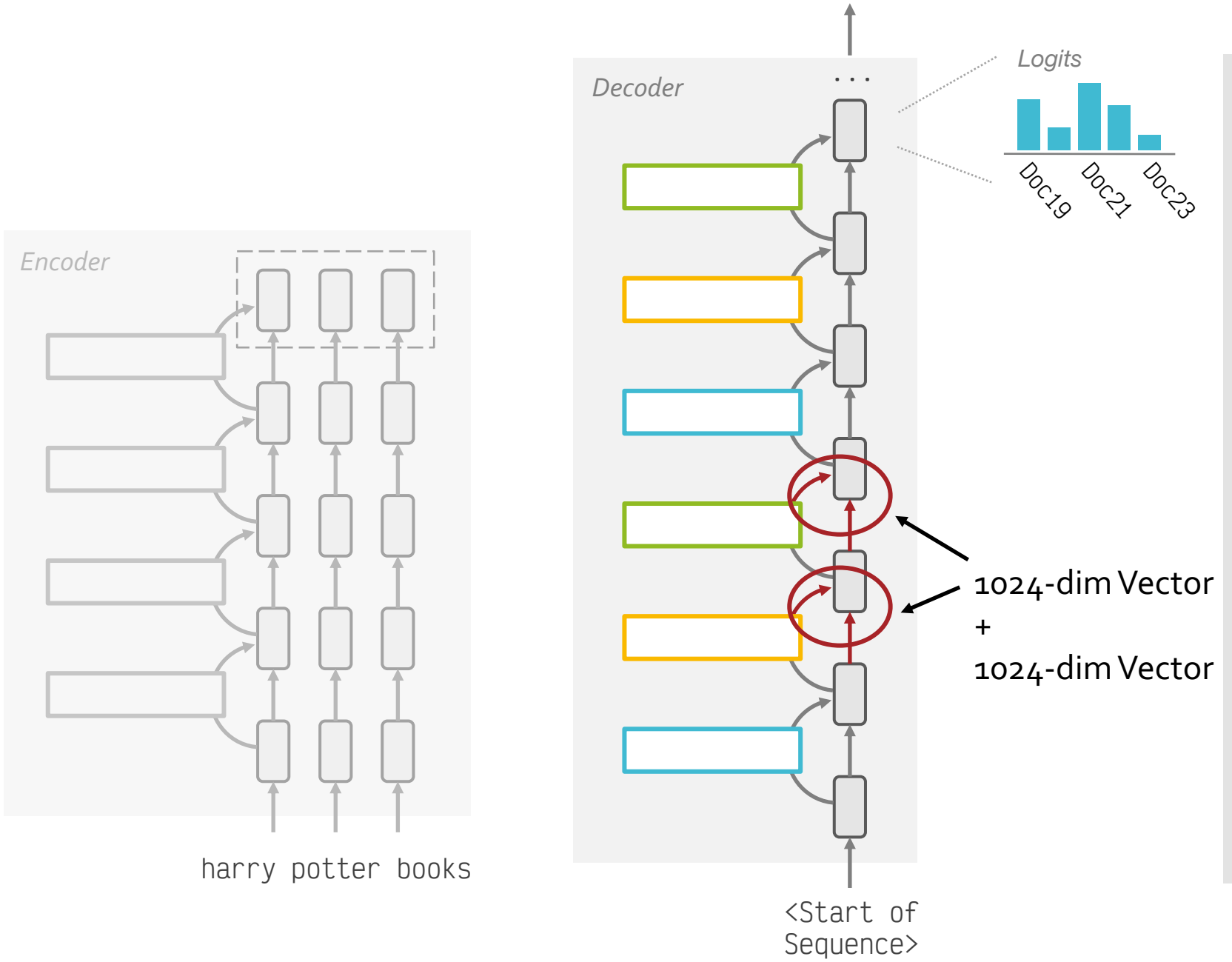
How does a GenIR Model look like?



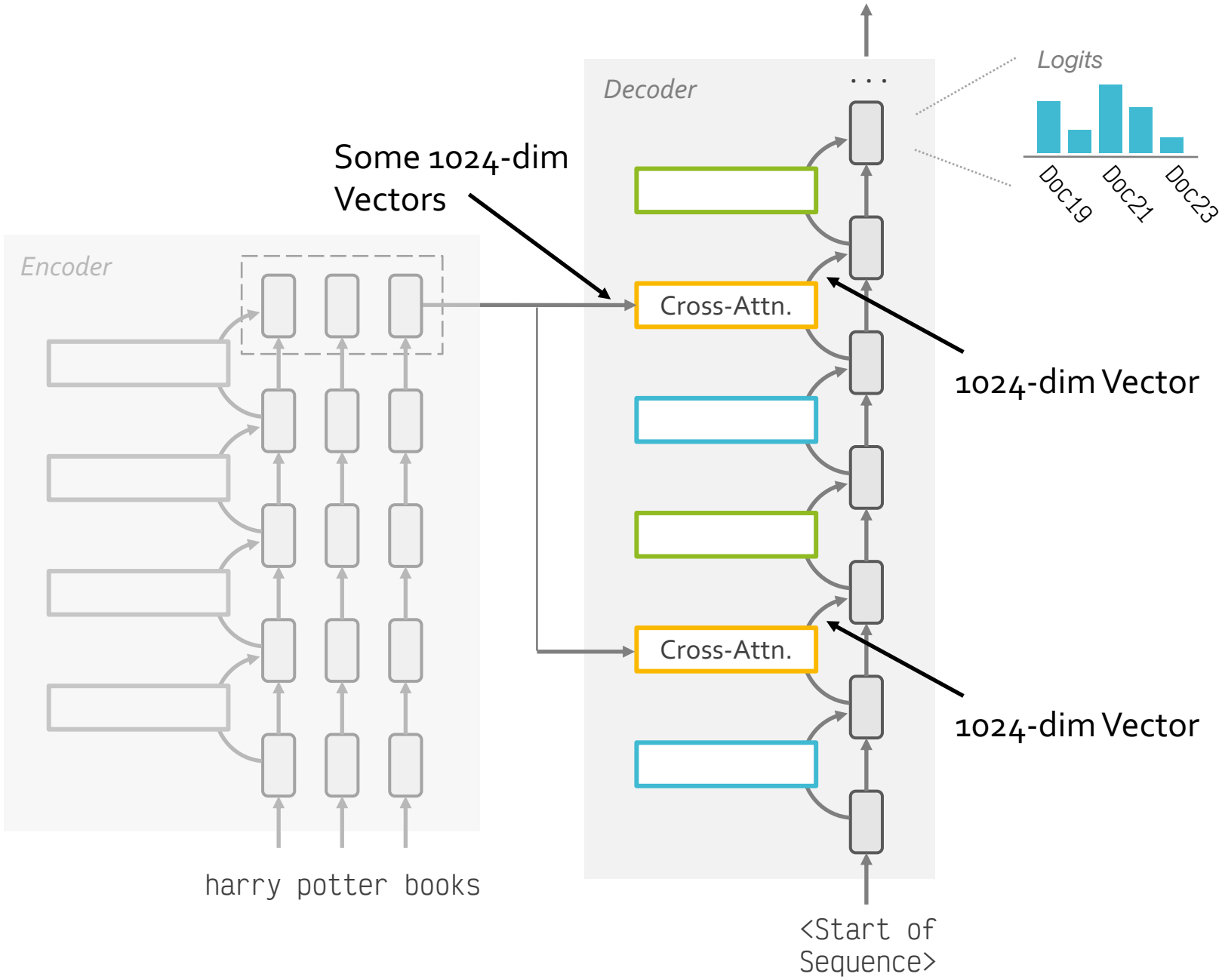
How does a GenIR Model look like?



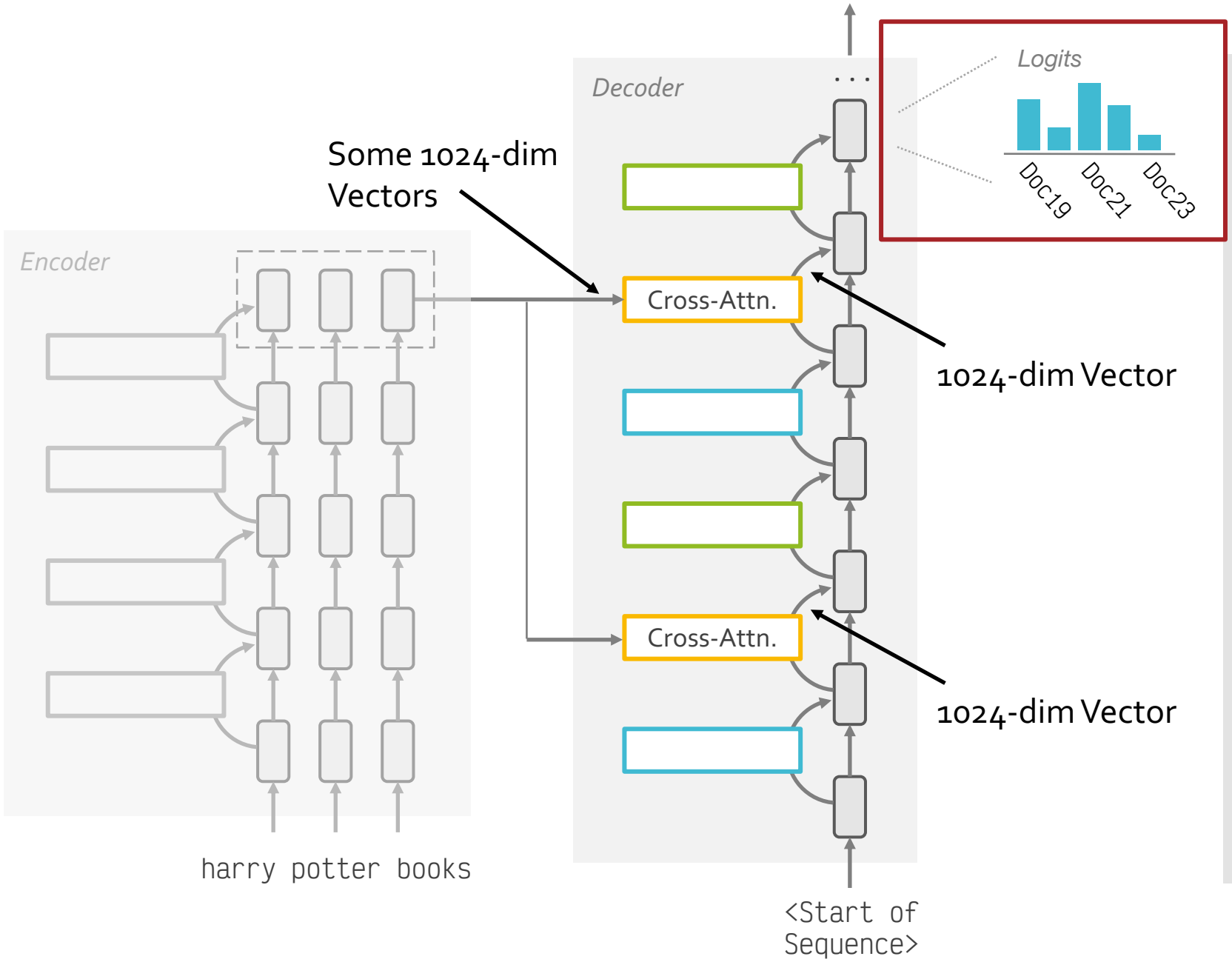
How does a GenIR Model look like?



How does a GenIR Model look like?

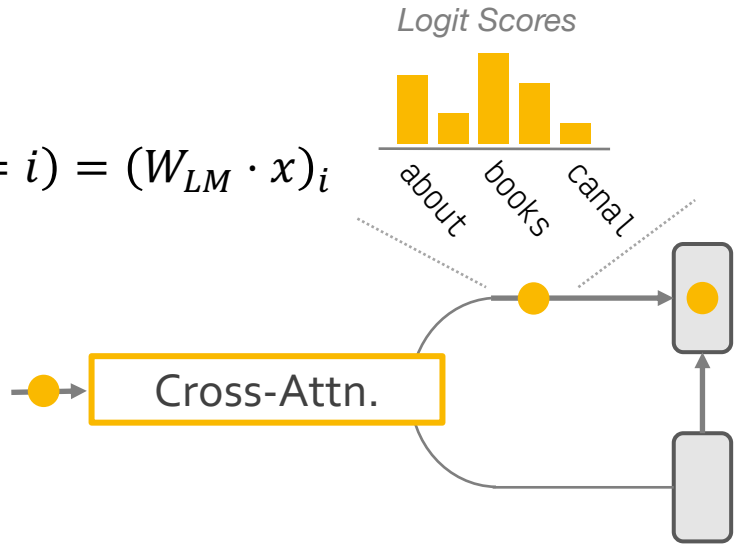


How does a GenIR Model look like?



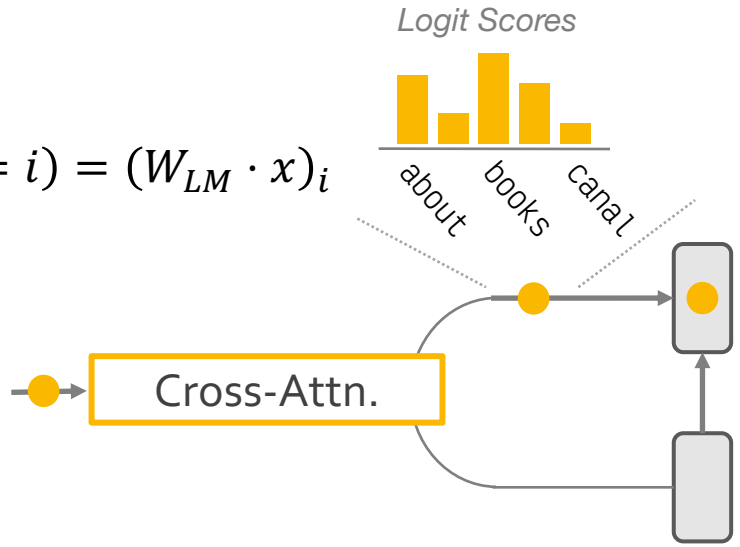
What does Cross-Attention write?

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



What does Cross-Attention write?

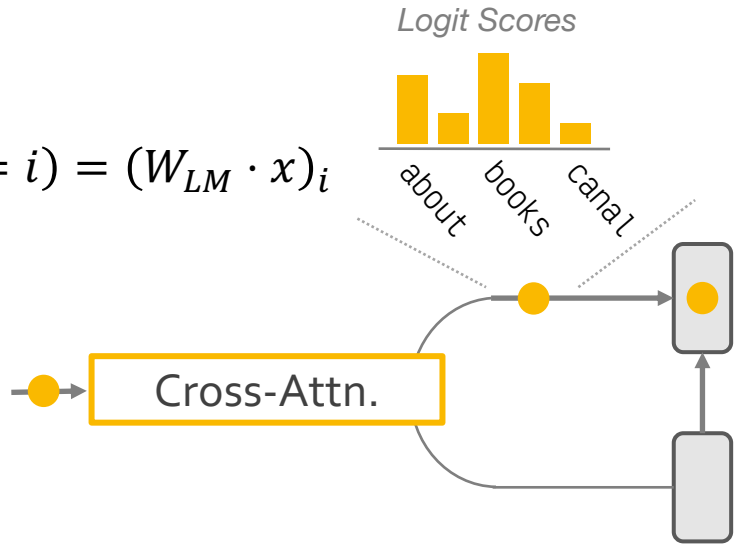
$$\mathrm{L}(token = i) = (W_{LM} \cdot x)_i$$



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?

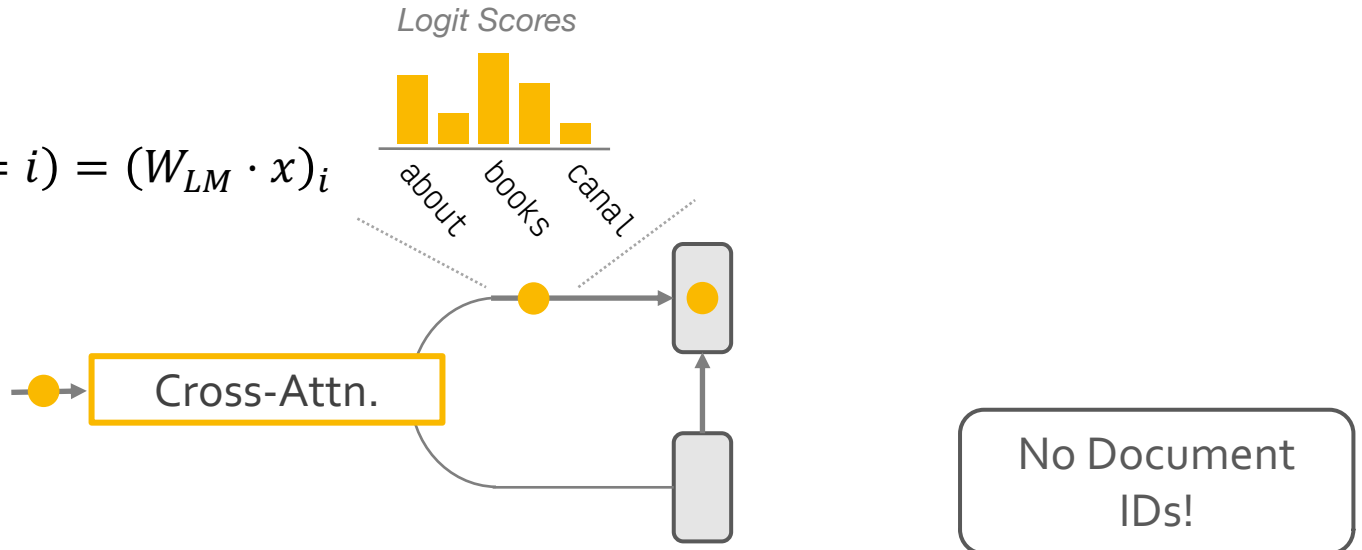
$$\mathrm{L}(token = i) = (W_{LM} \cdot x)_i$$



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

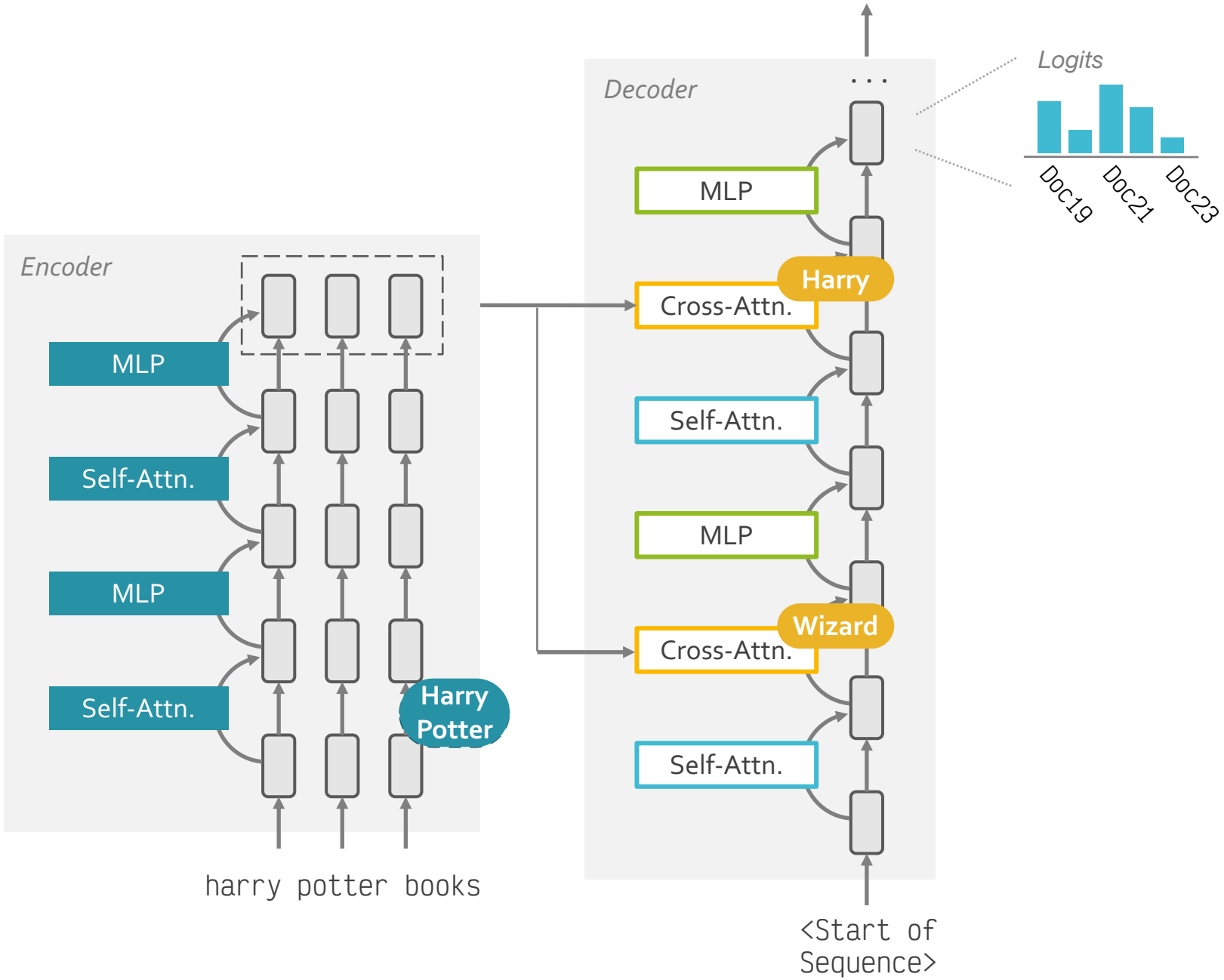
What does Cross-Attention write?

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

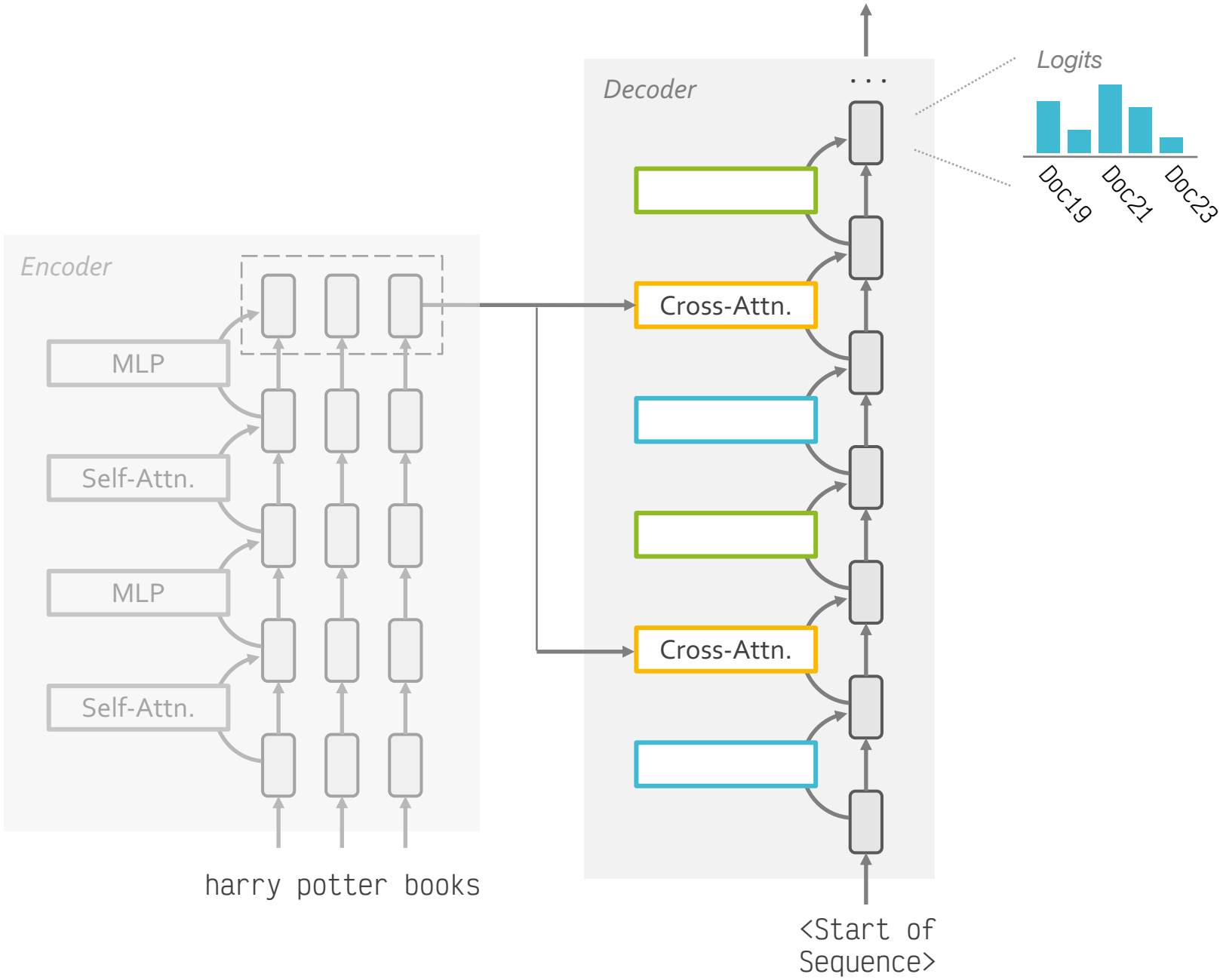


QUERY	LAYER	HEAD	TOP 5 TOKENS
who wrote the harry potter books	Layer 14	Head 2	about, written, about, tailored, privire
	Layer 16	Head 1	books, ouvrage, books, authors, book
who won the football championship in 2002	Layer 16	Head 1	year, YEAR, Year, year, jahr
	Layer 16	Head 13	football, Football, fotbal,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

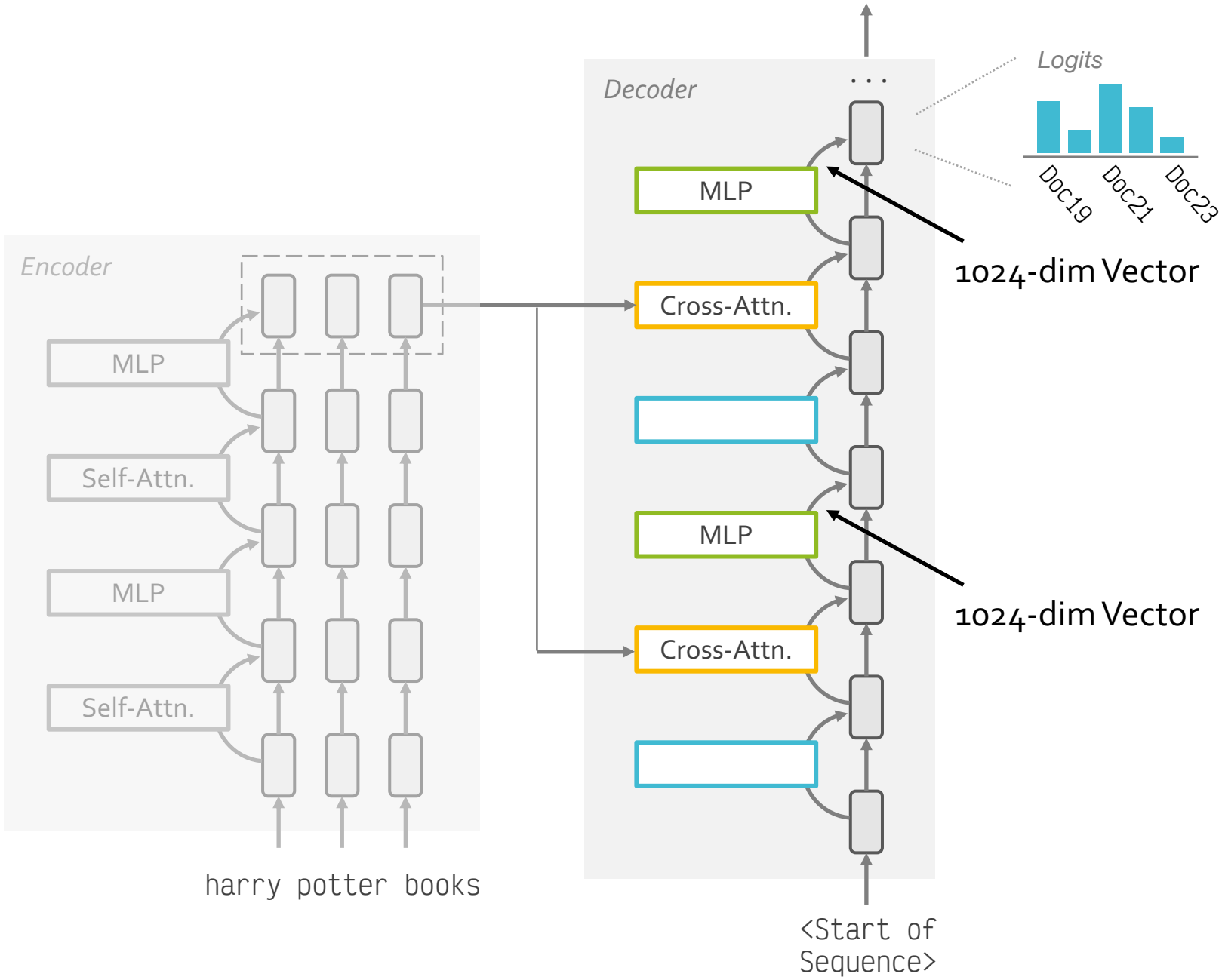
The Process so far



The GenIR Model

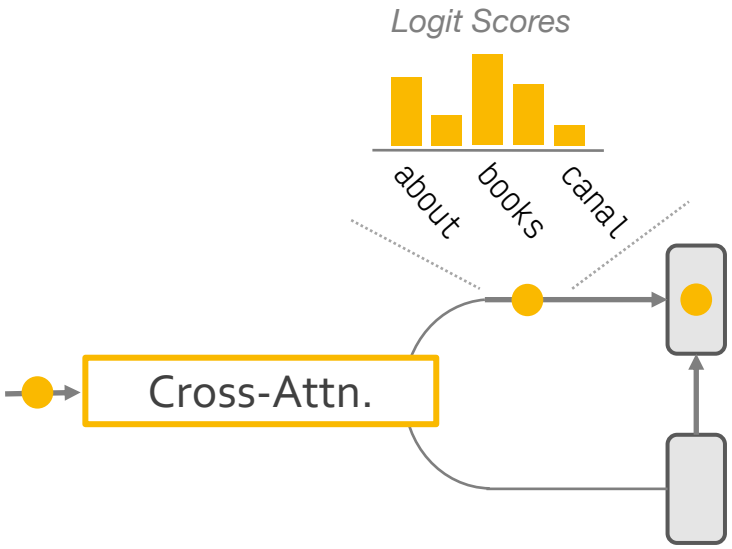


The GenIR Model



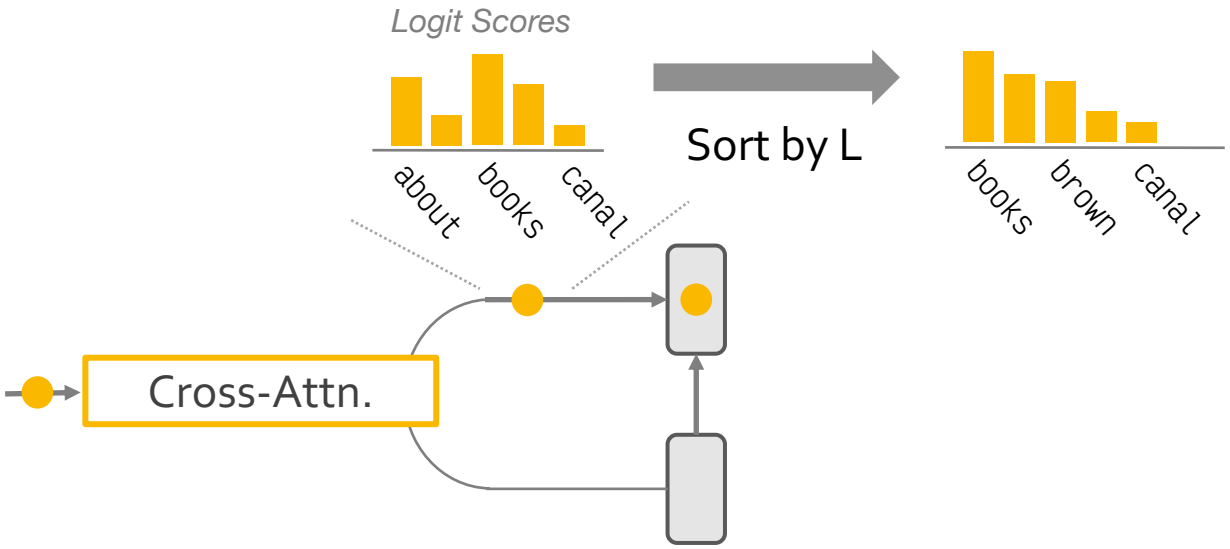
Component Output

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



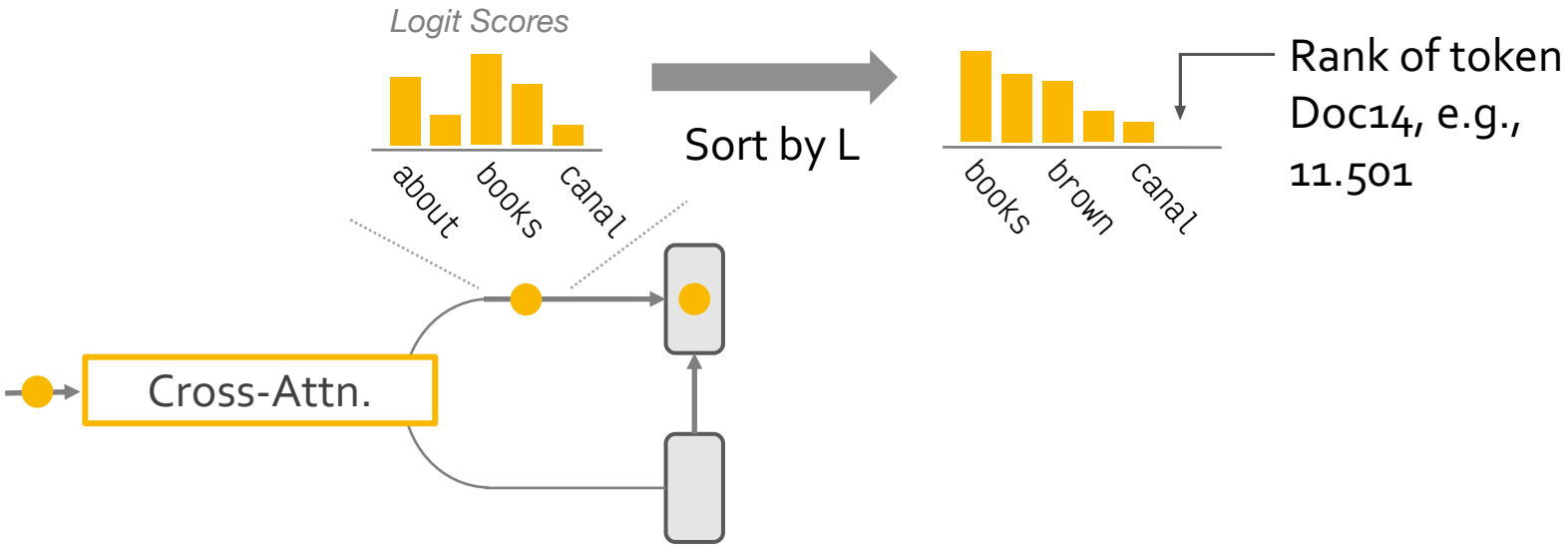
Component Output

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



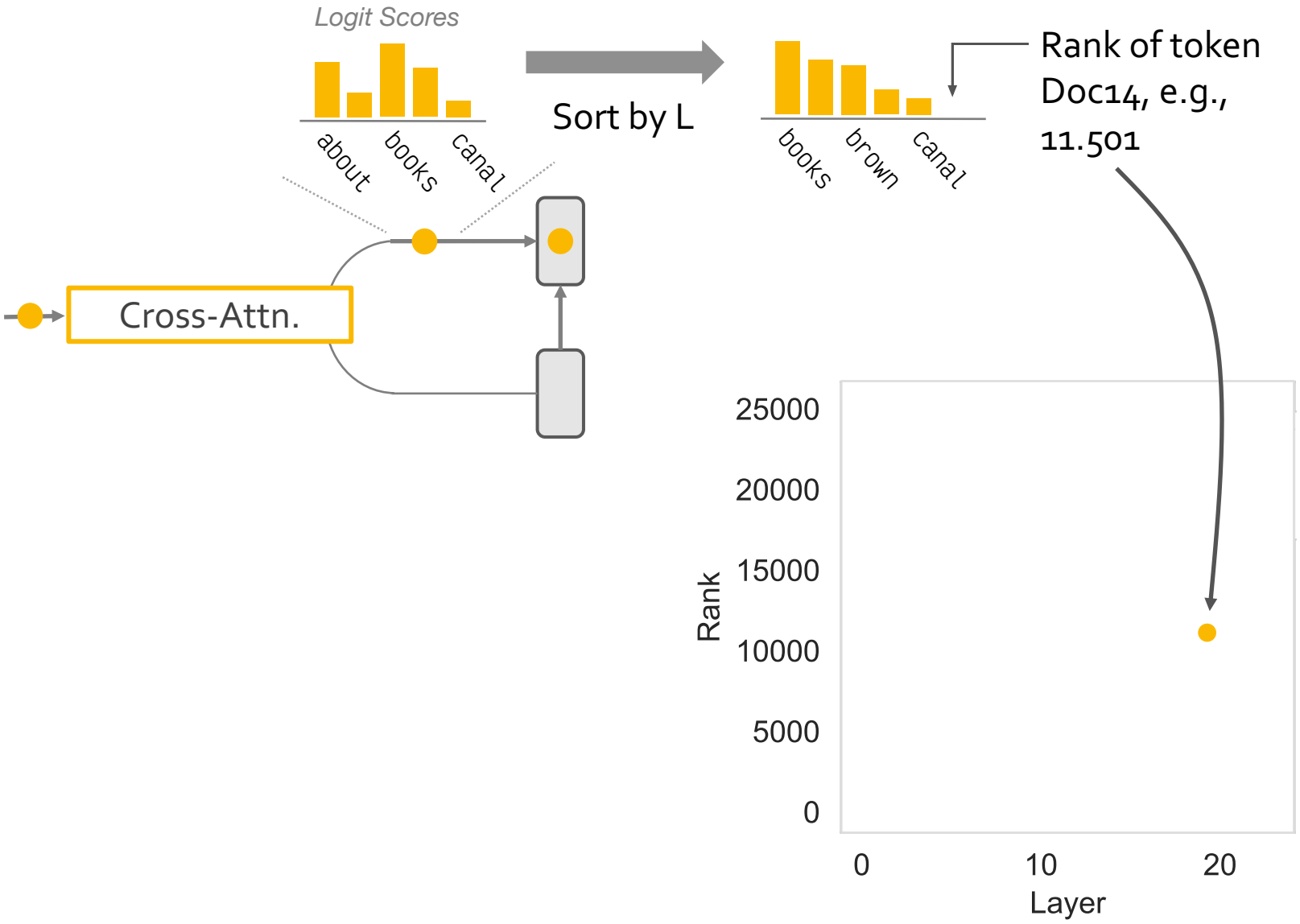
Component Output

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



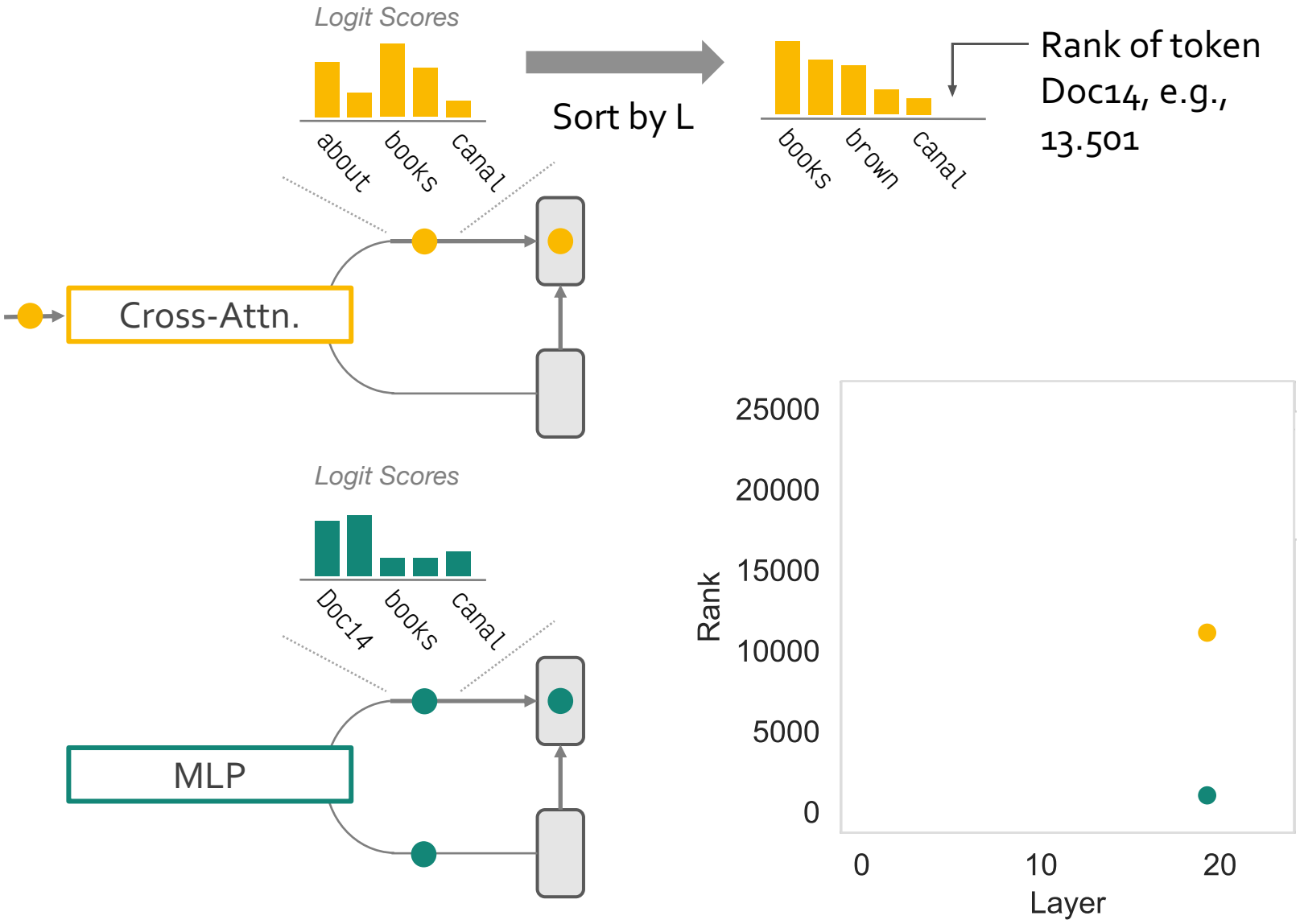
Component Output

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



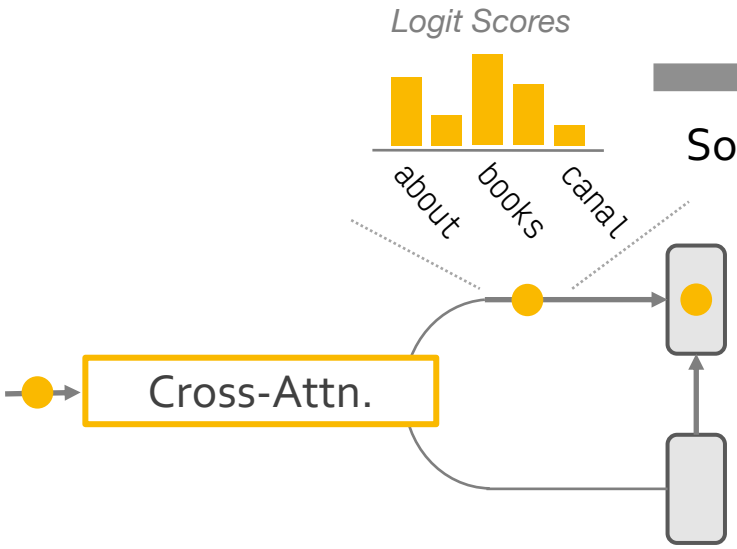
Component Output

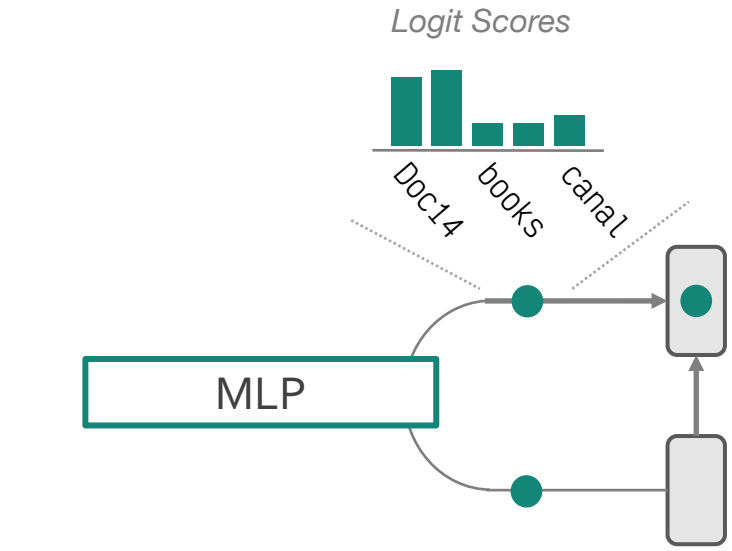
$$\mathrm{L}(token = i) = (W_{LM} \cdot x)_i$$



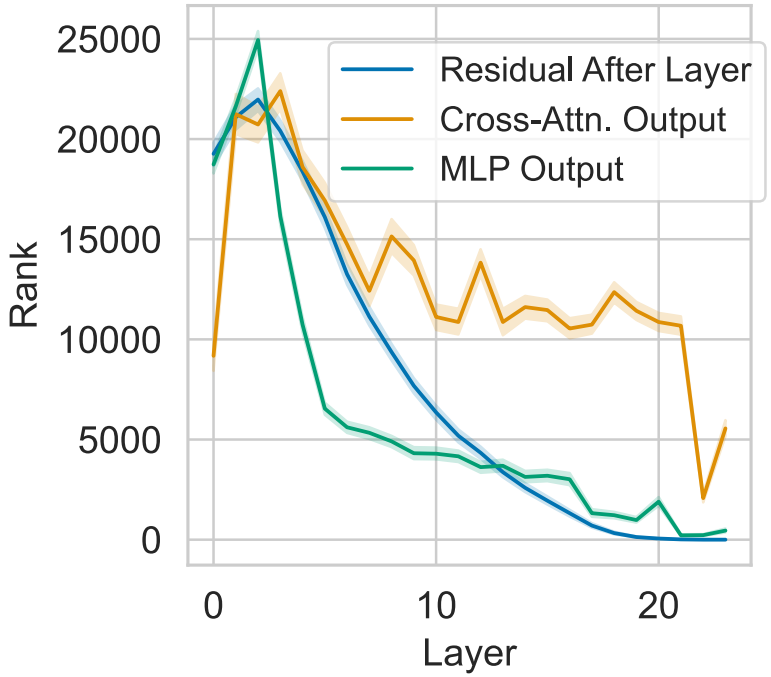
Component Output

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

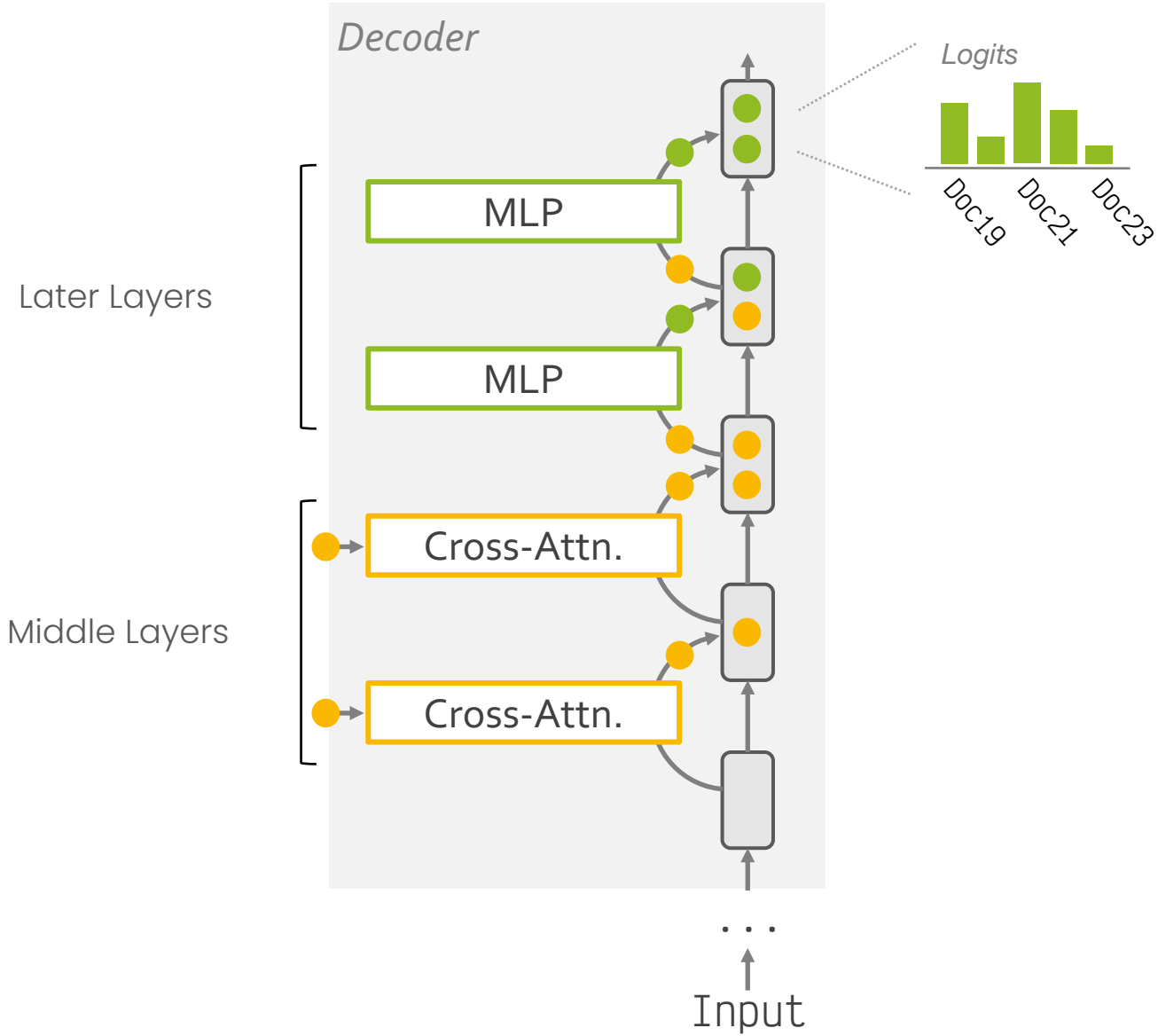




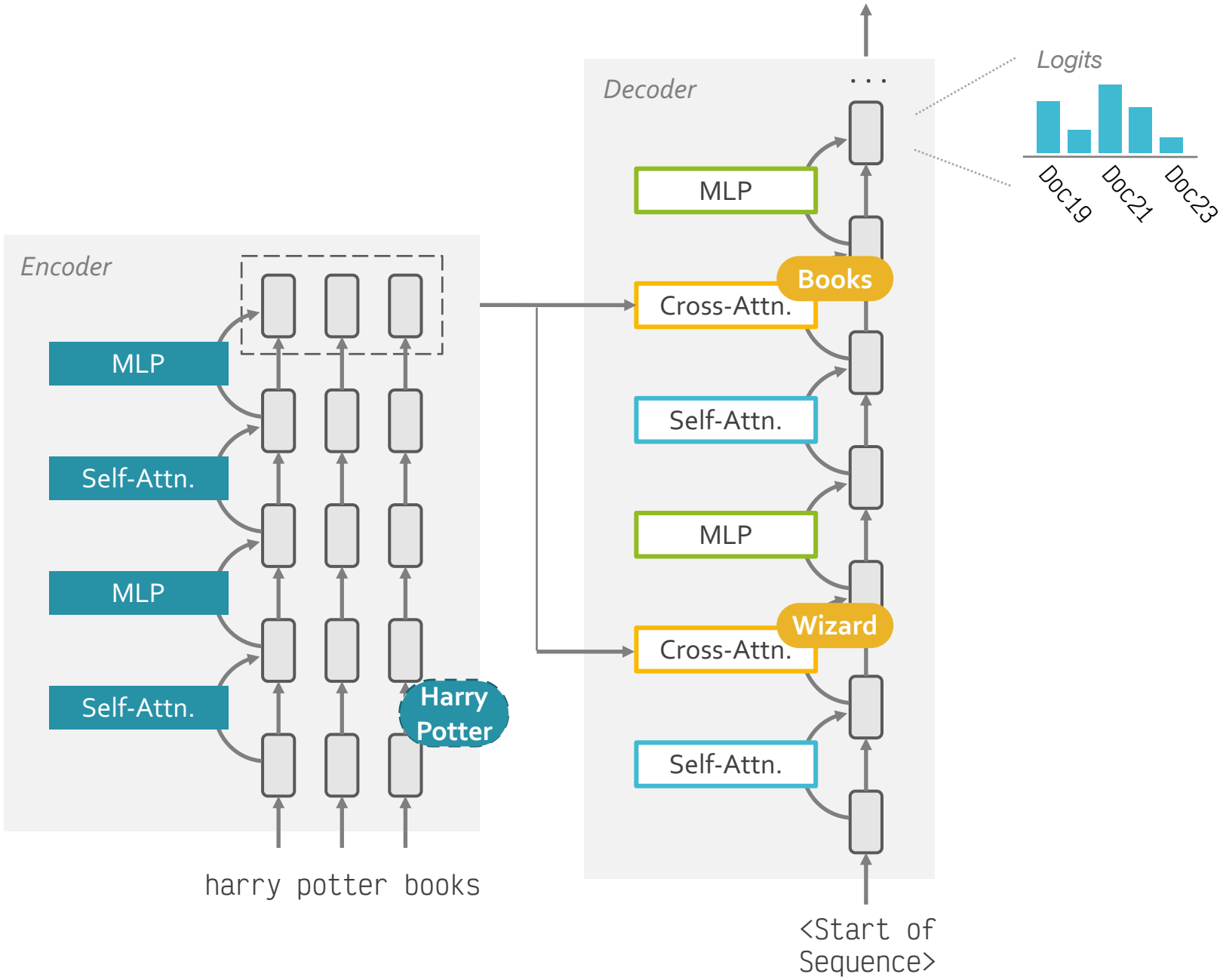
Rank of token Doc14, e.g., 13.501



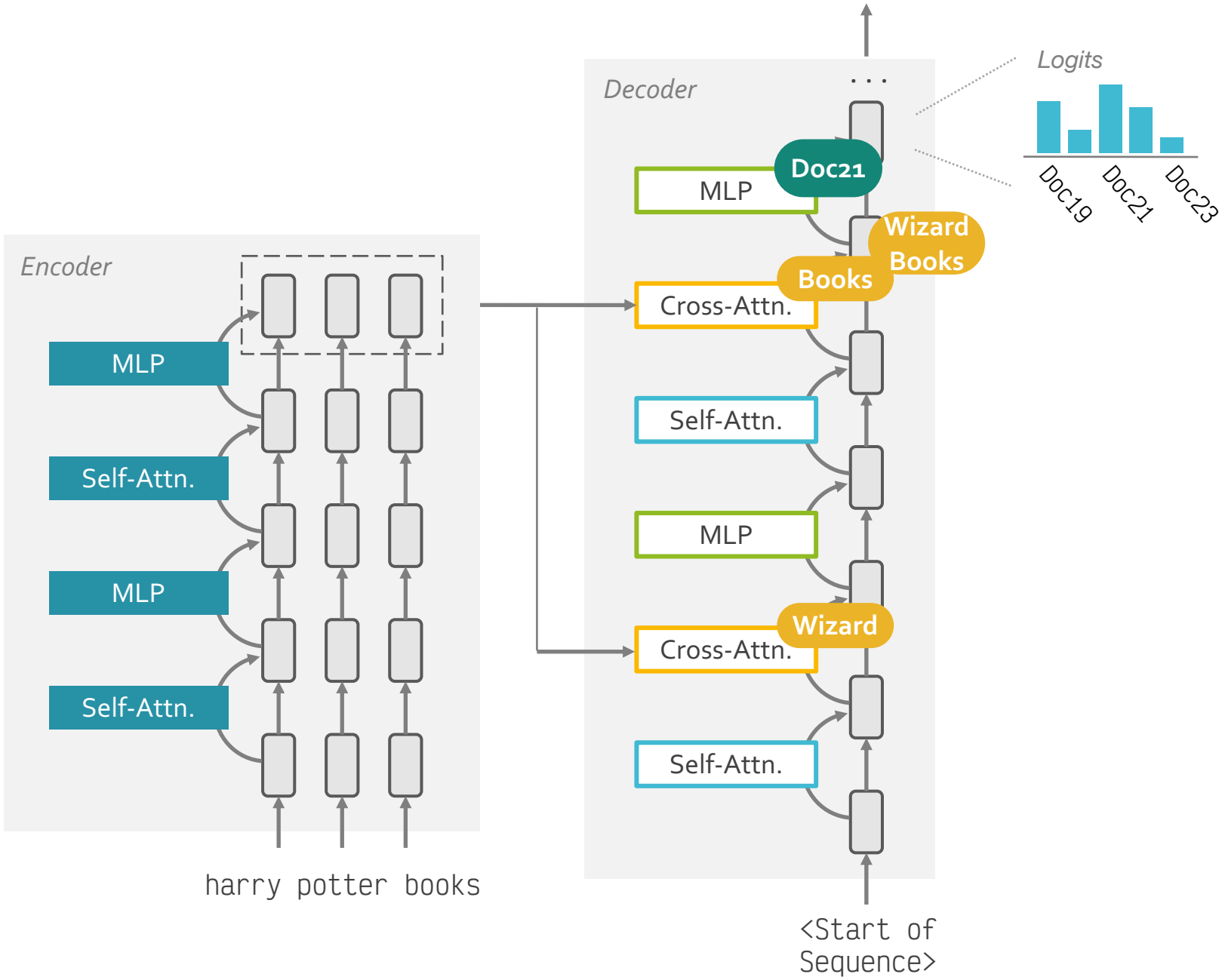
Cross-Attention writes, what MLPs later read.



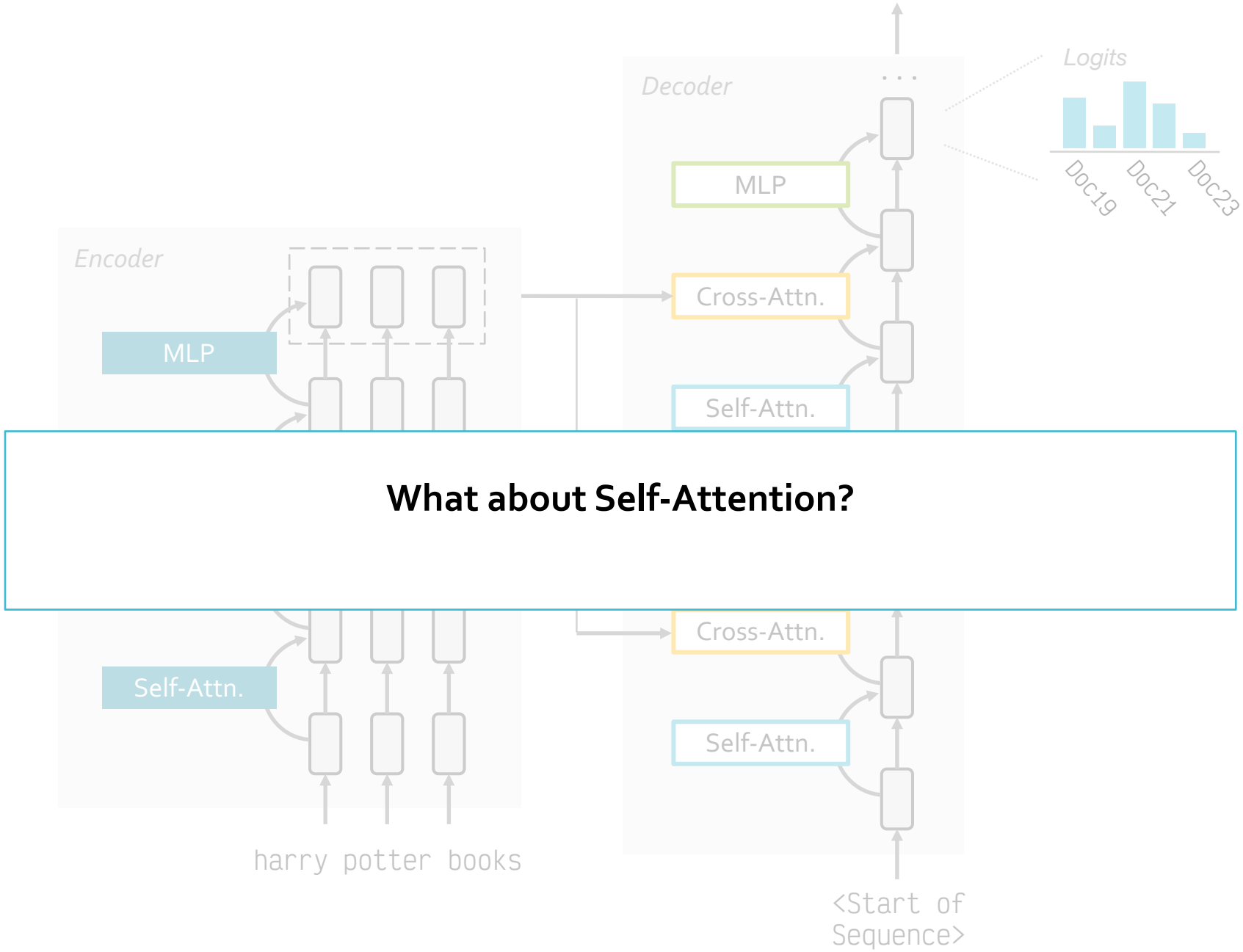
The whole picture (?)



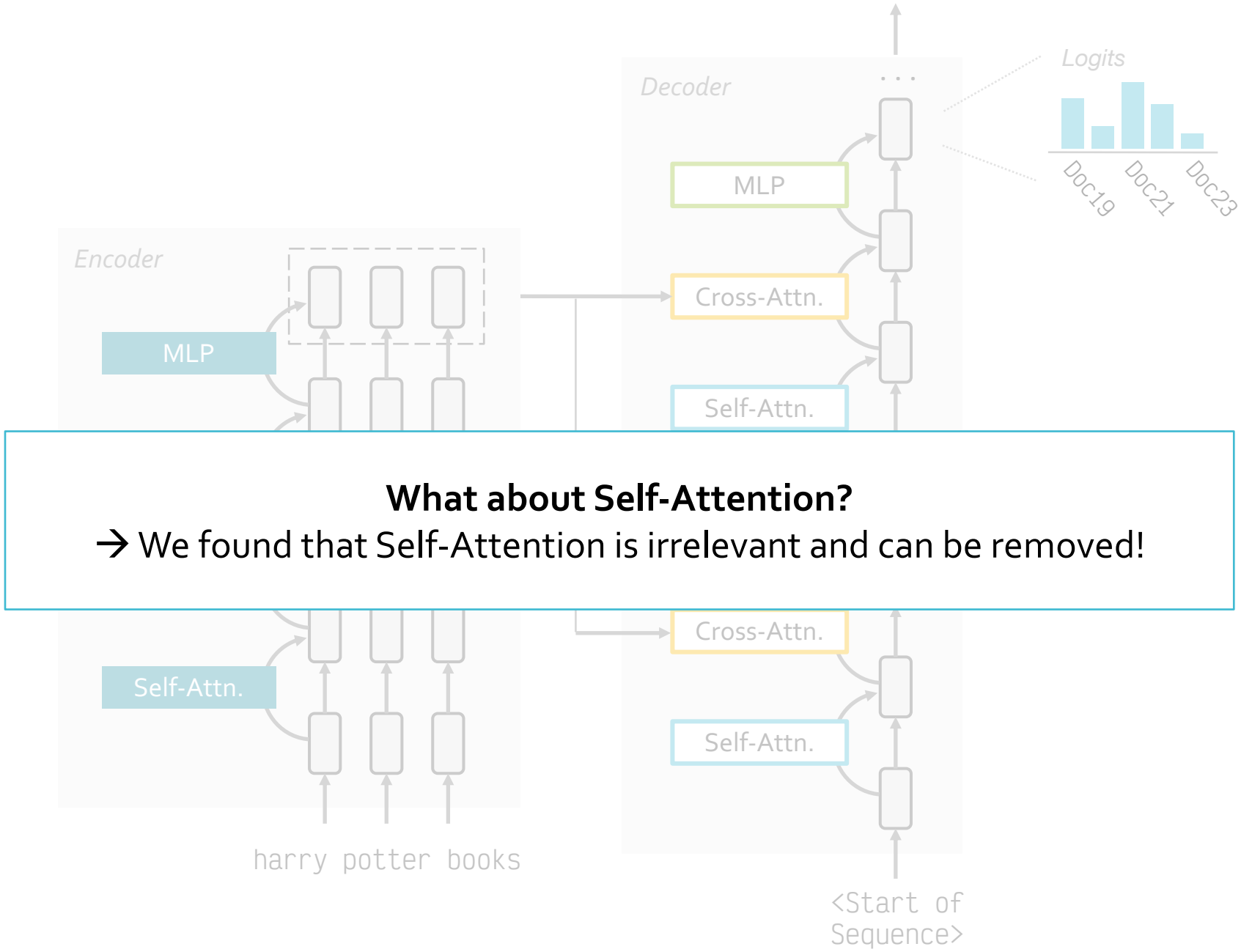
The whole picture (?)



The whole picture (?)



The whole picture (?)

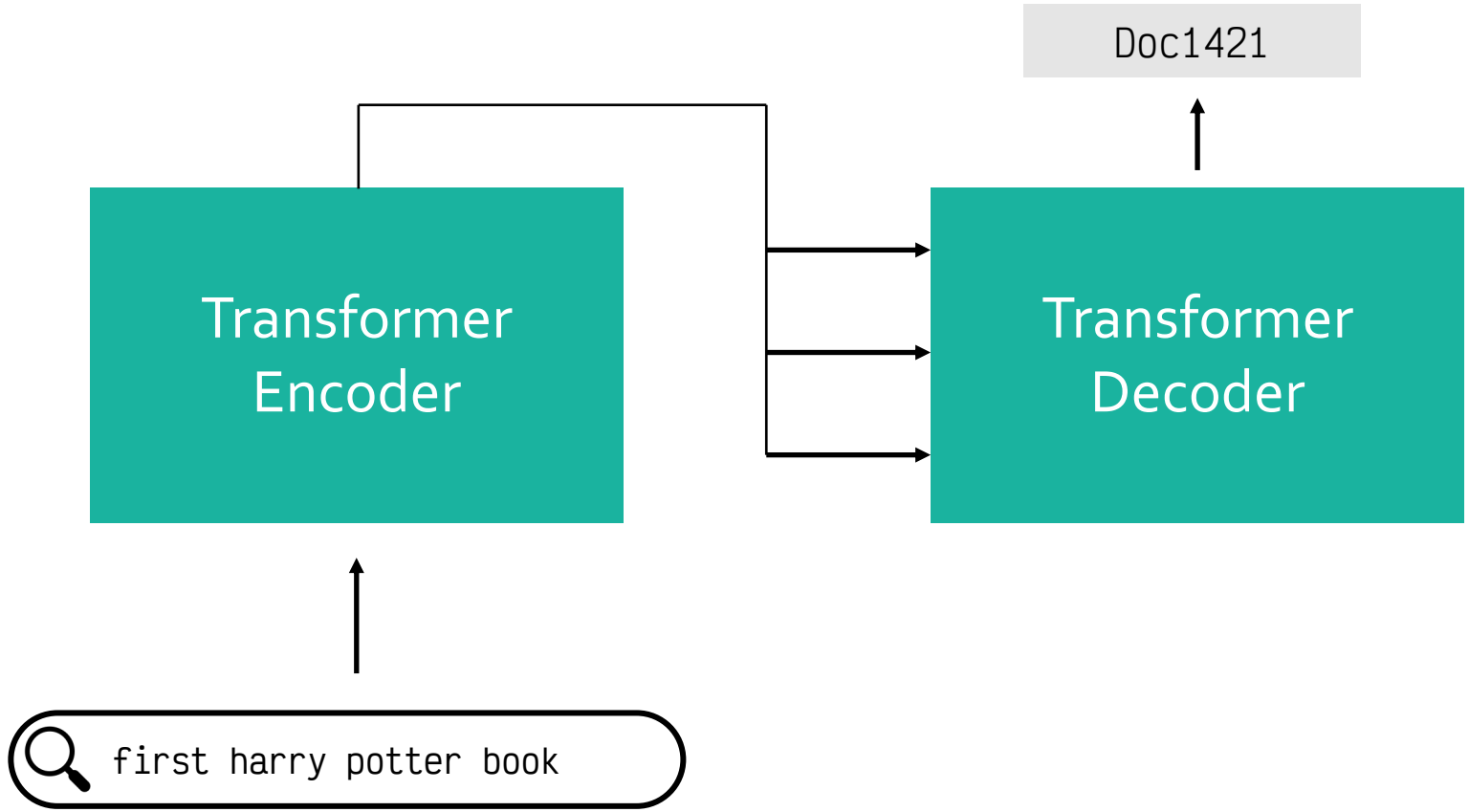


What about Self-Attention?

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

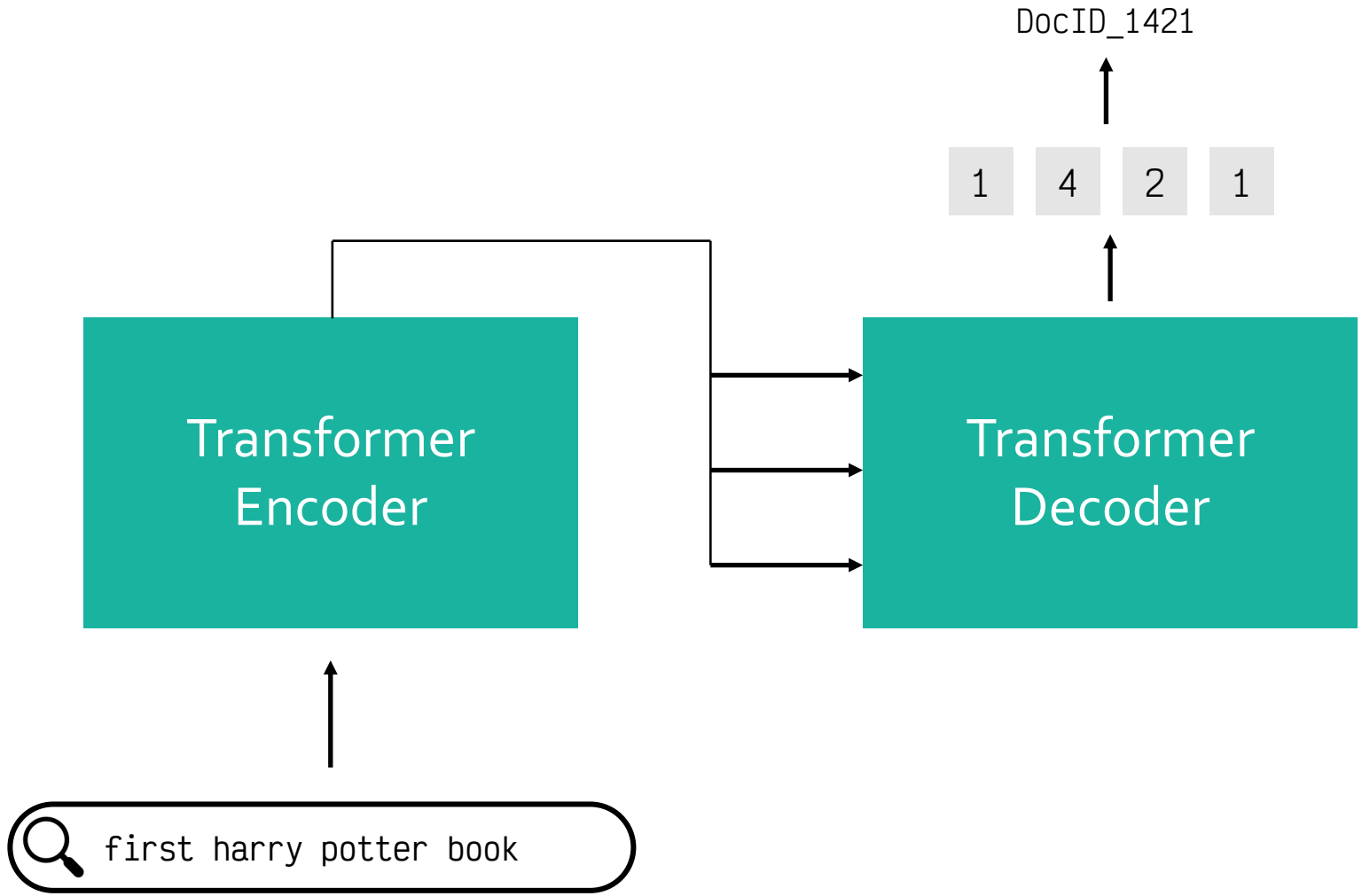
The Role of Self-Attention in Generative IR

Single-Token vs. Multi-Token



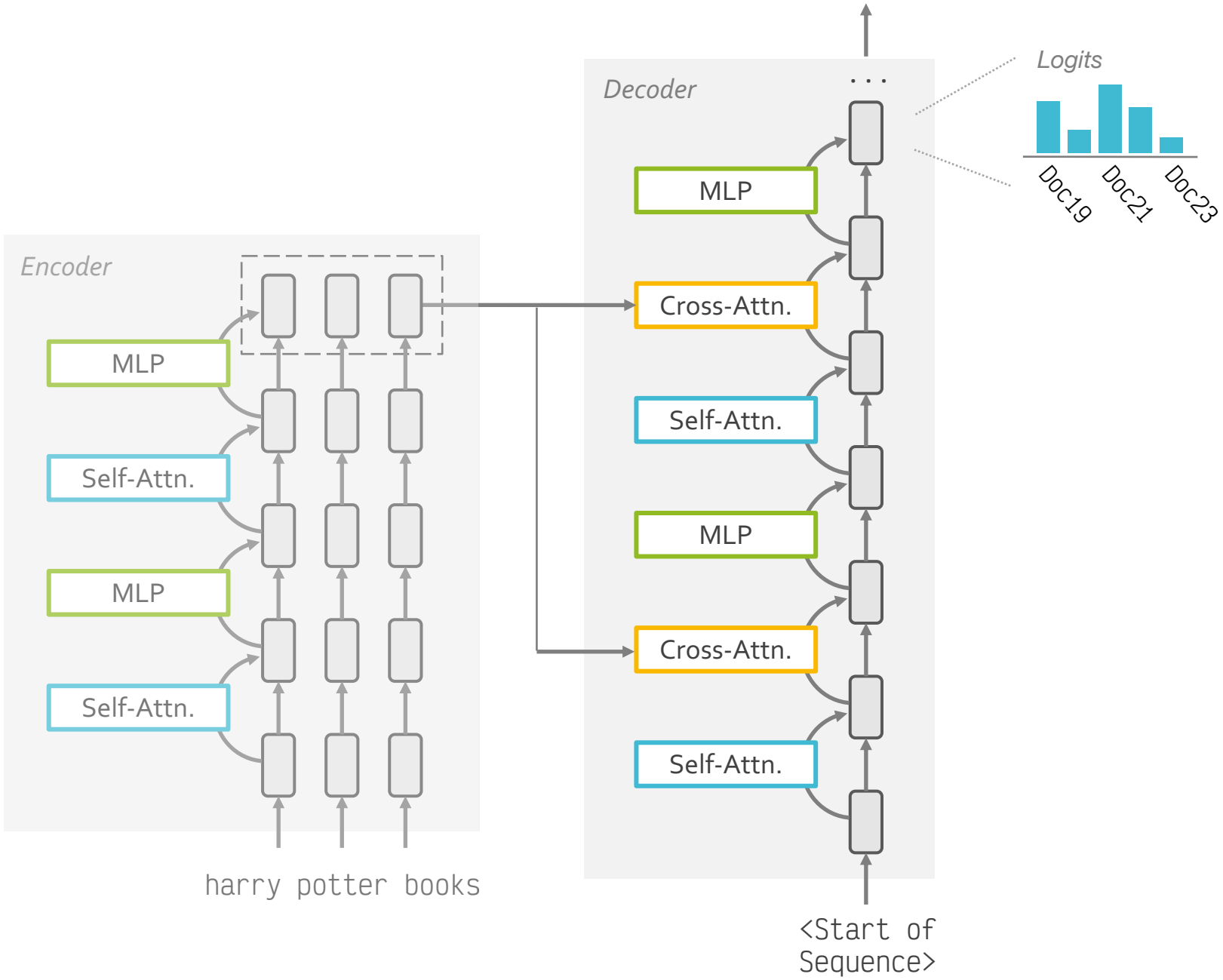
Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

Single-Token vs. Multi-Token

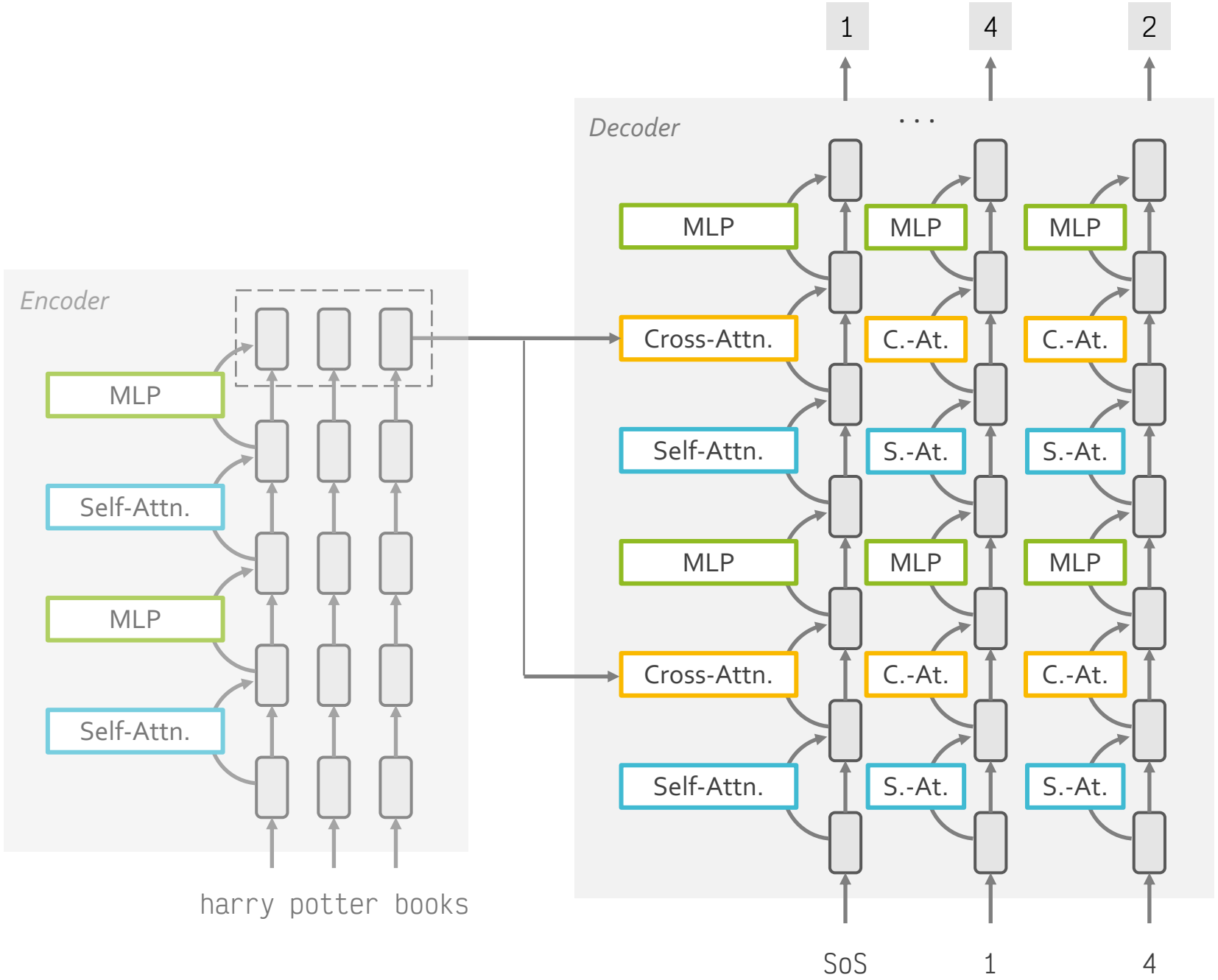


Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, *35*, 21831-21843.

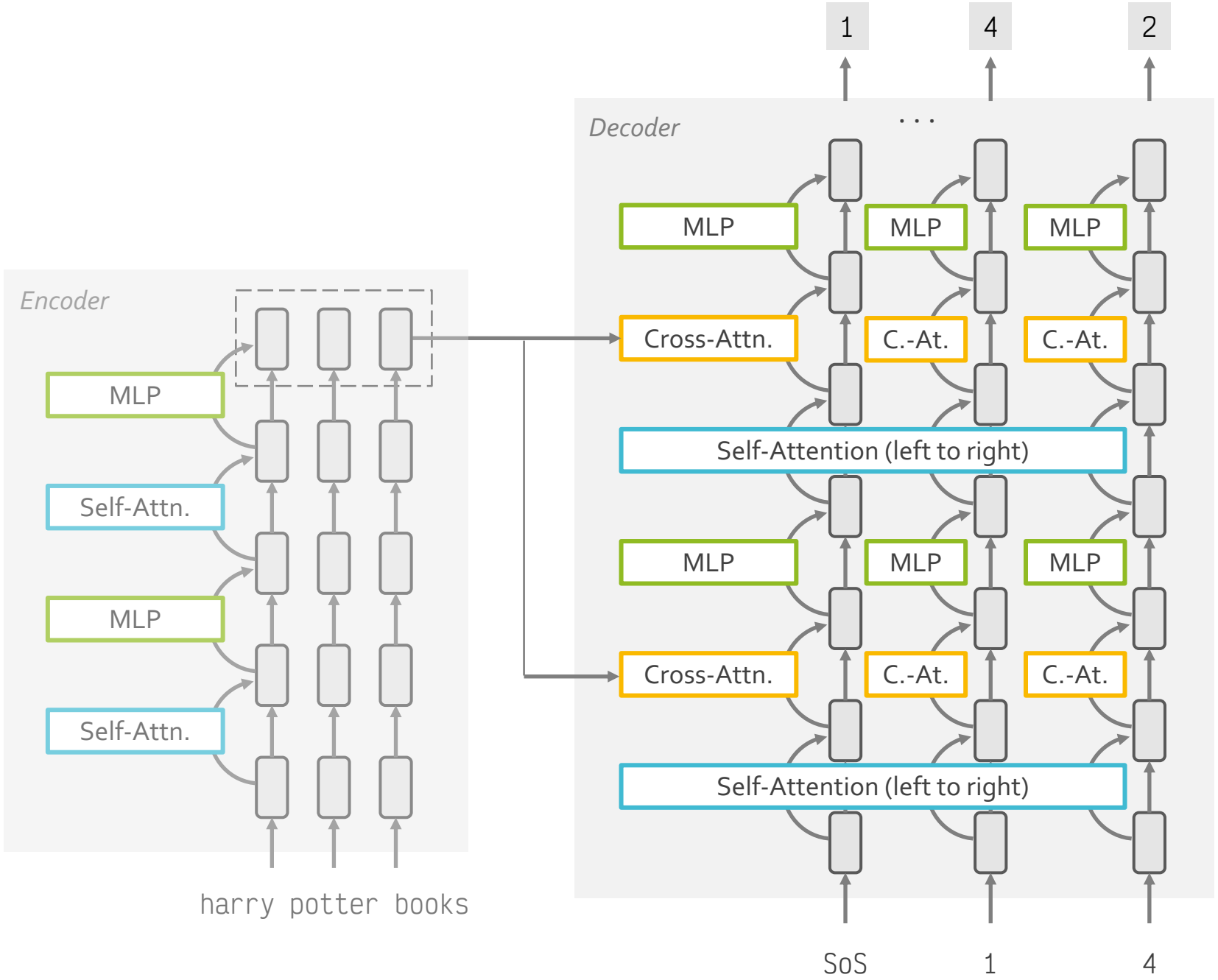
Single-Token Prediction



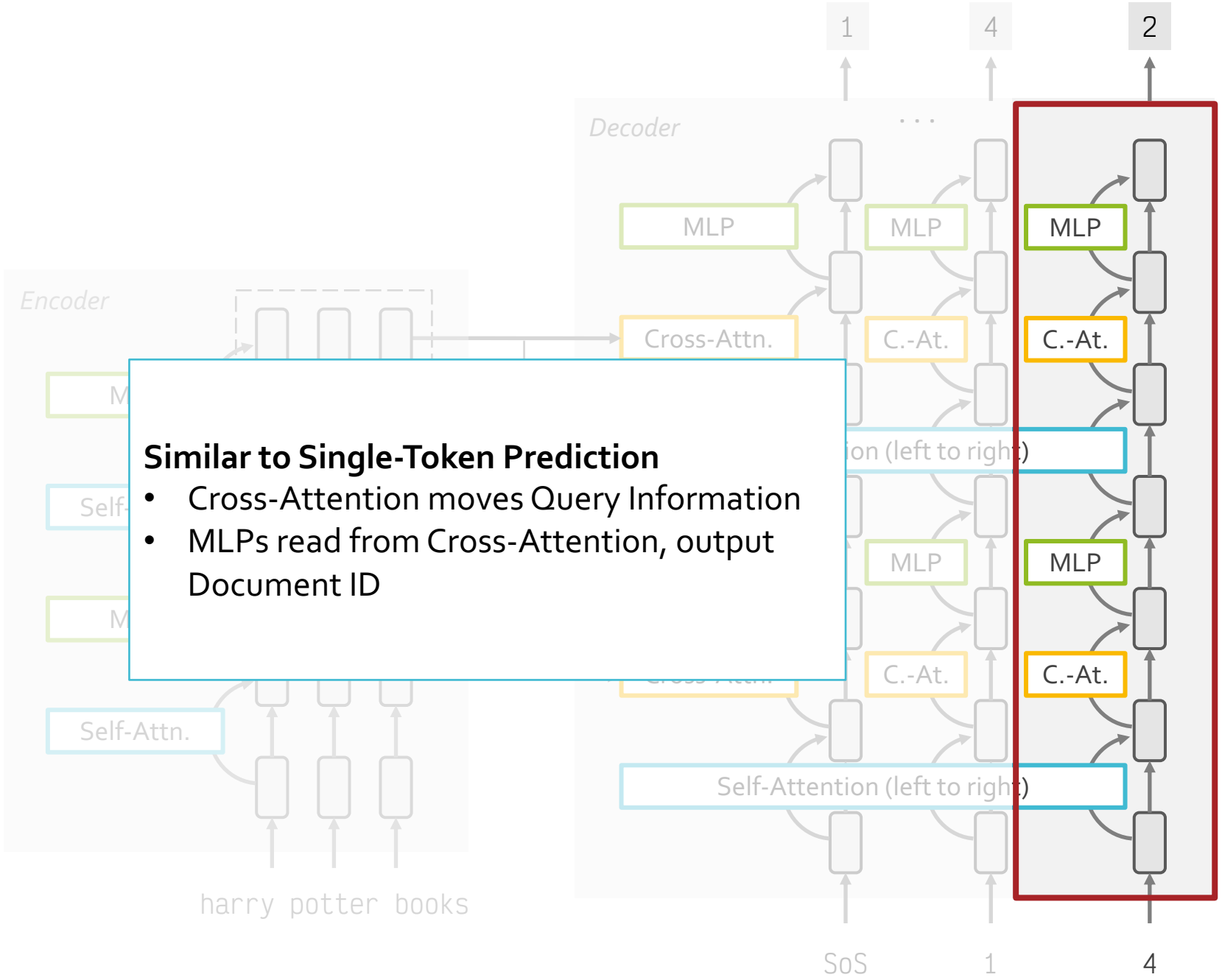
Multi-Token Prediction



Multi-Token Prediction



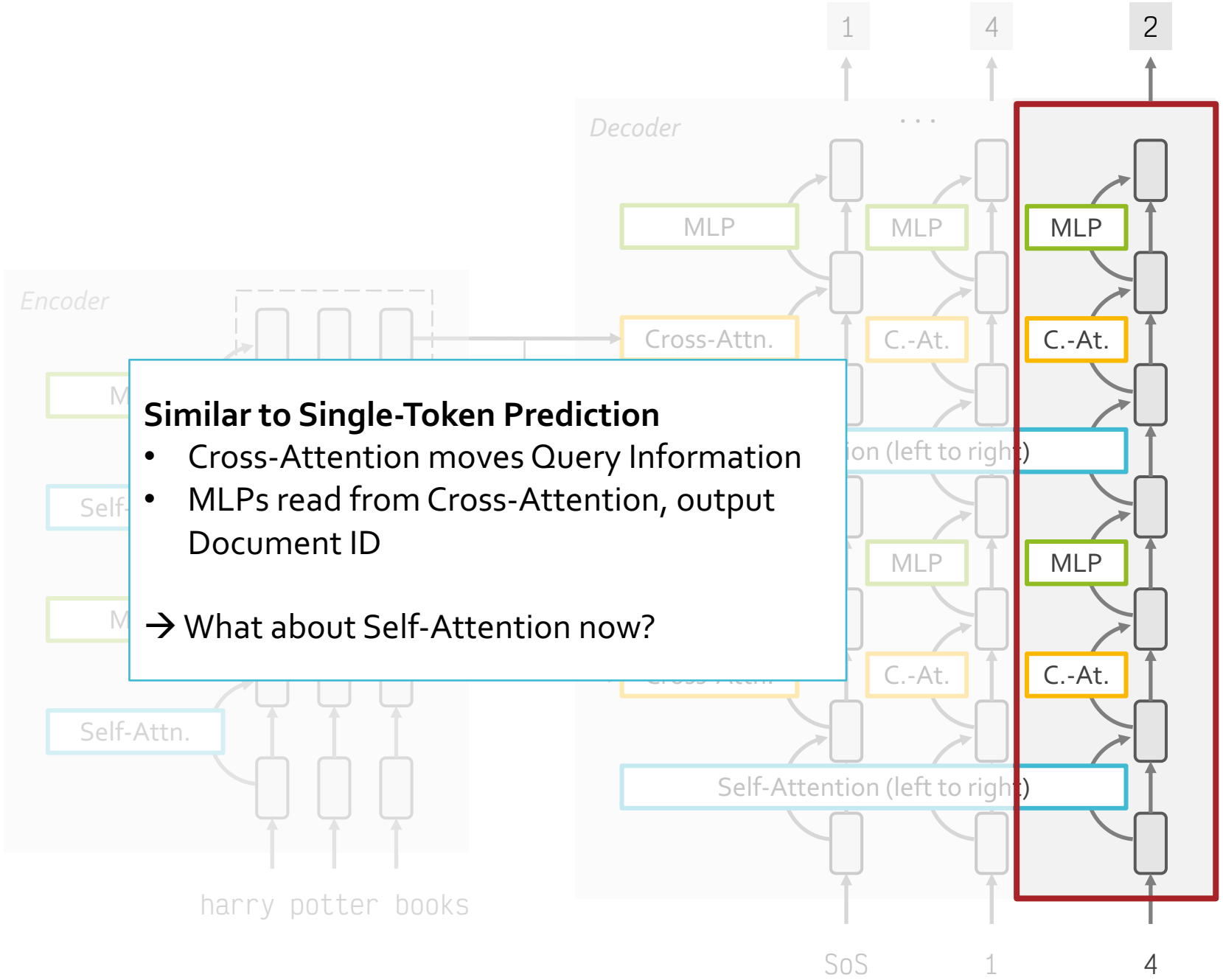
Multi-Token Prediction



Similar to Single-Token Prediction

- Cross-Attention moves Query Information
- MLPs read from Cross-Attention, output Document ID

Multi-Token Prediction

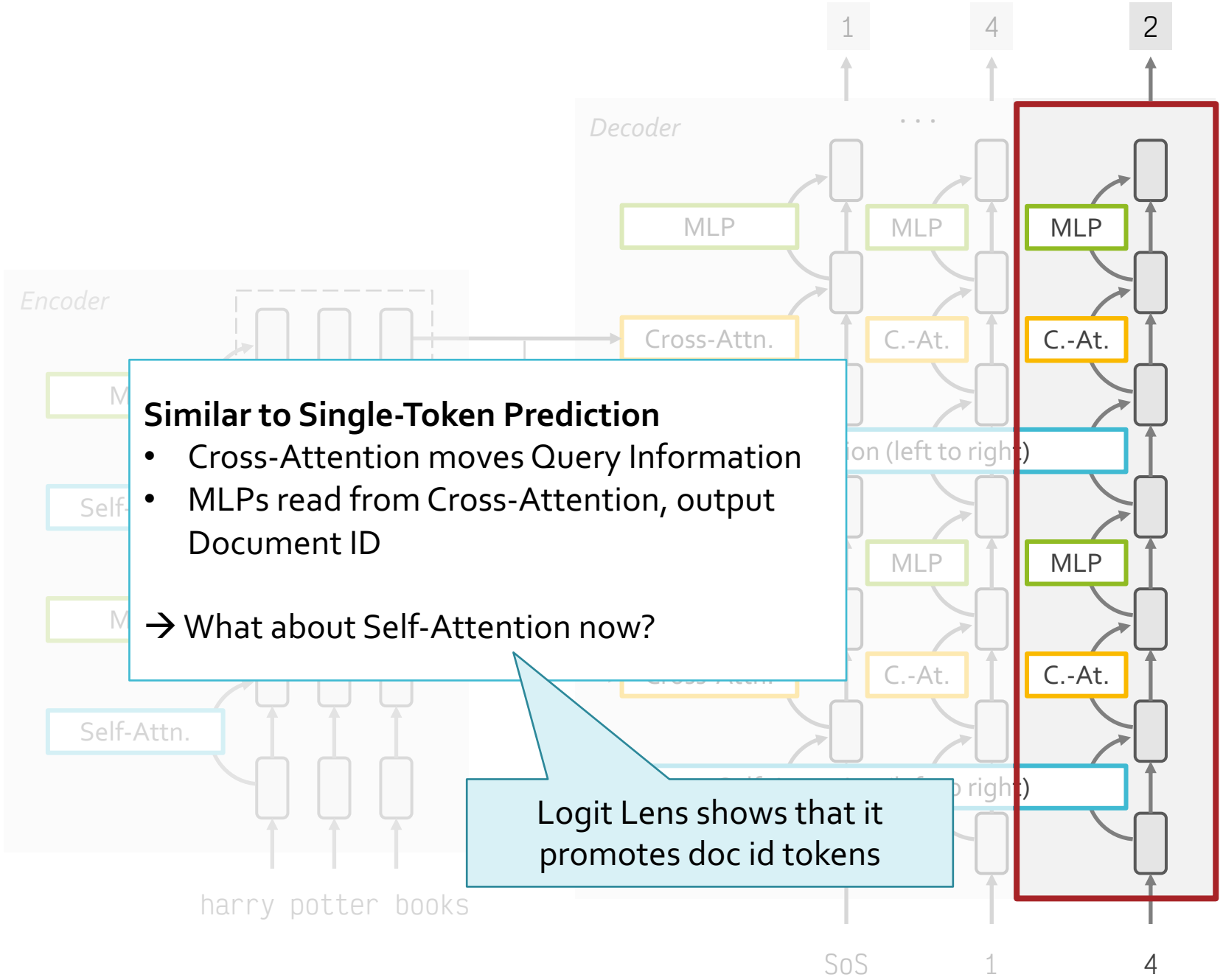


Similar to Single-Token Prediction

- Cross-Attention moves Query Information
- MLPs read from Cross-Attention, output Document ID

→ What about Self-Attention now?

Multi-Token Prediction



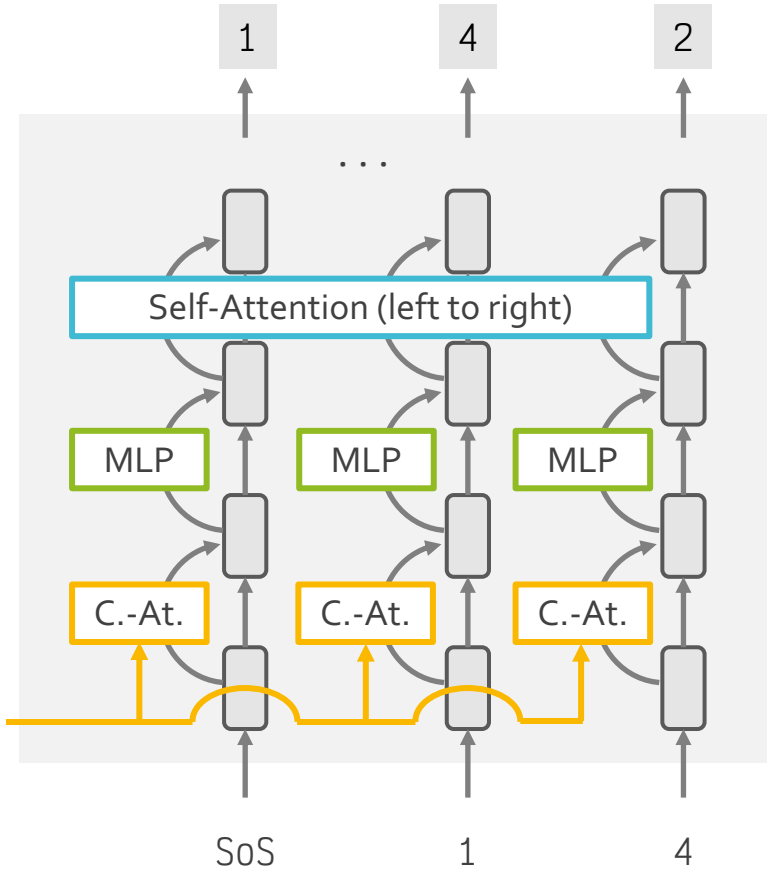
Similar to Single-Token Prediction

- Cross-Attention moves Query Information
- MLPs read from Cross-Attention, output Document ID

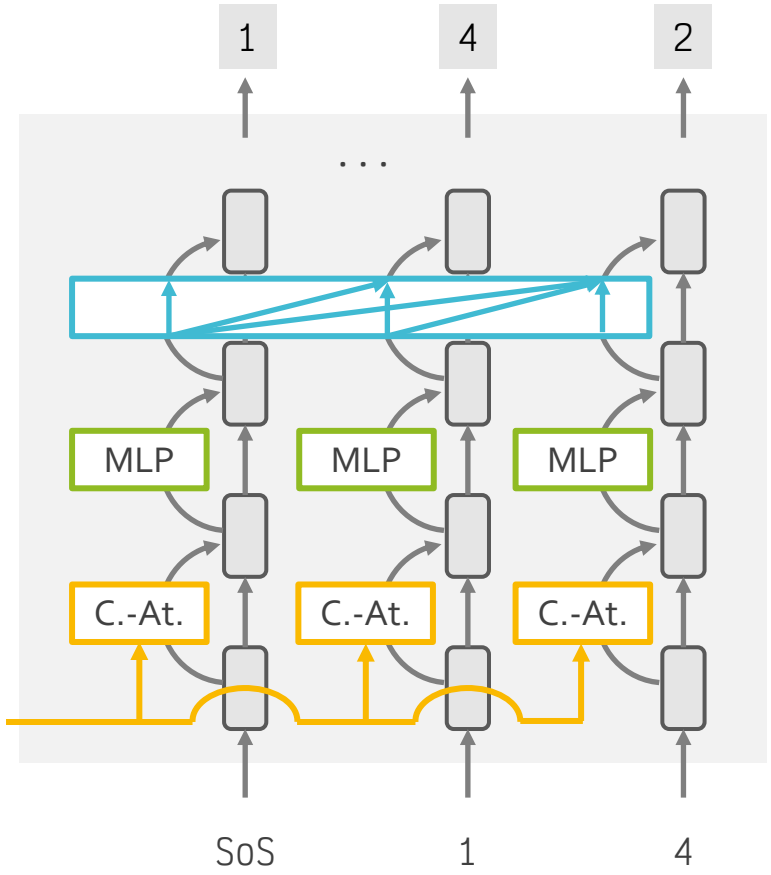
→ What about Self-Attention now?

Logit Lens shows that it promotes doc id tokens

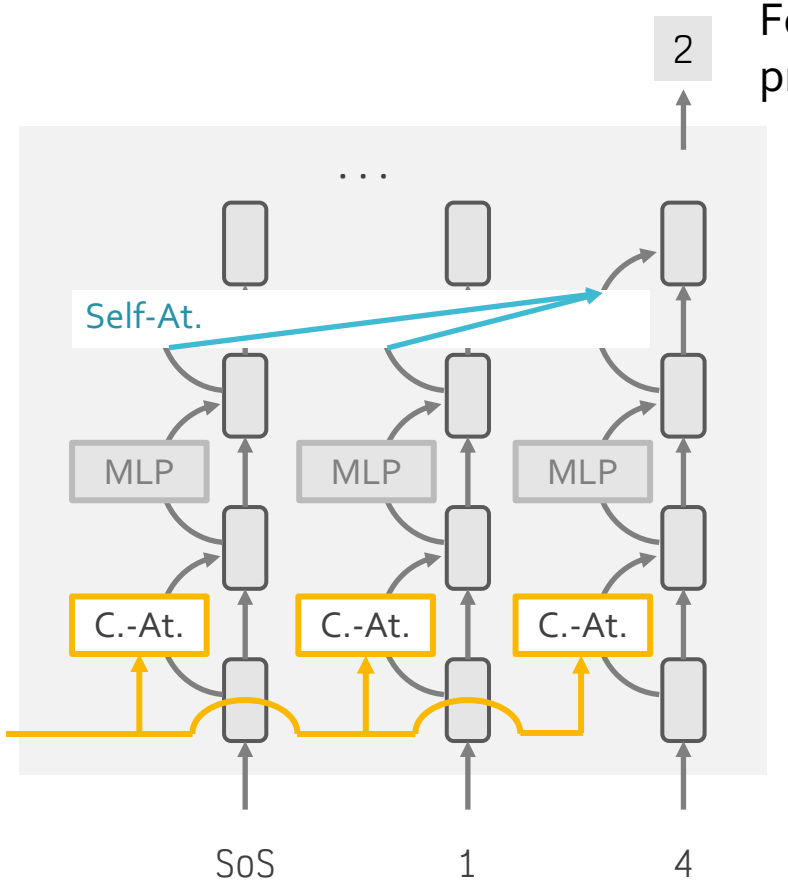
Information Movement through Self-Attention



Information Movement through Self-Attention

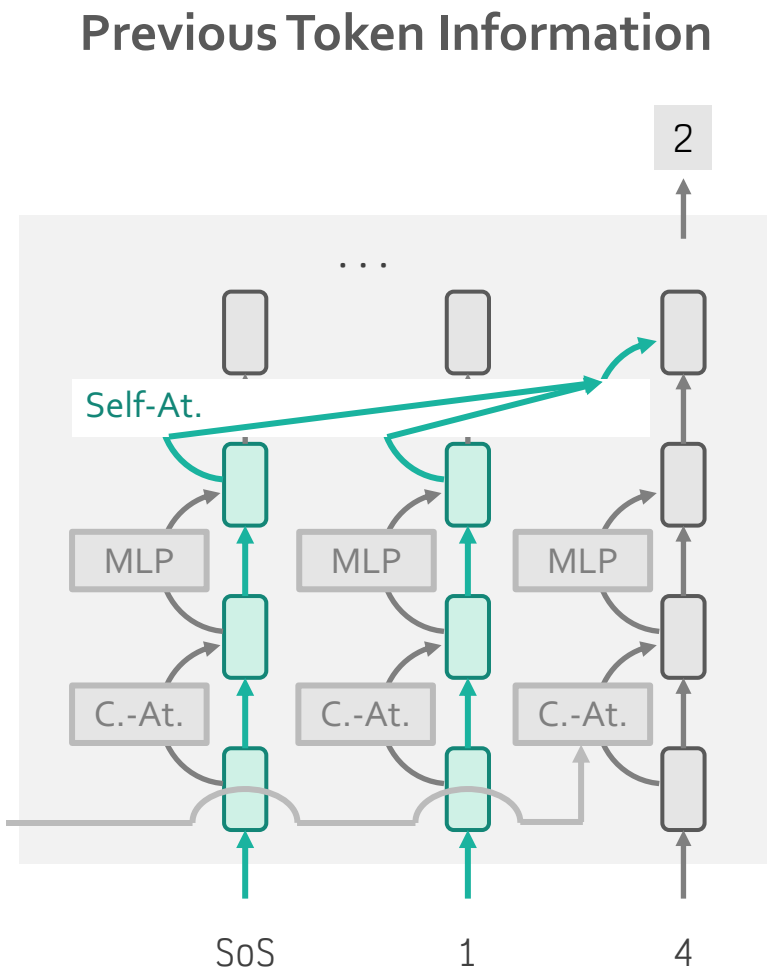


Information Movement through Self-Attention

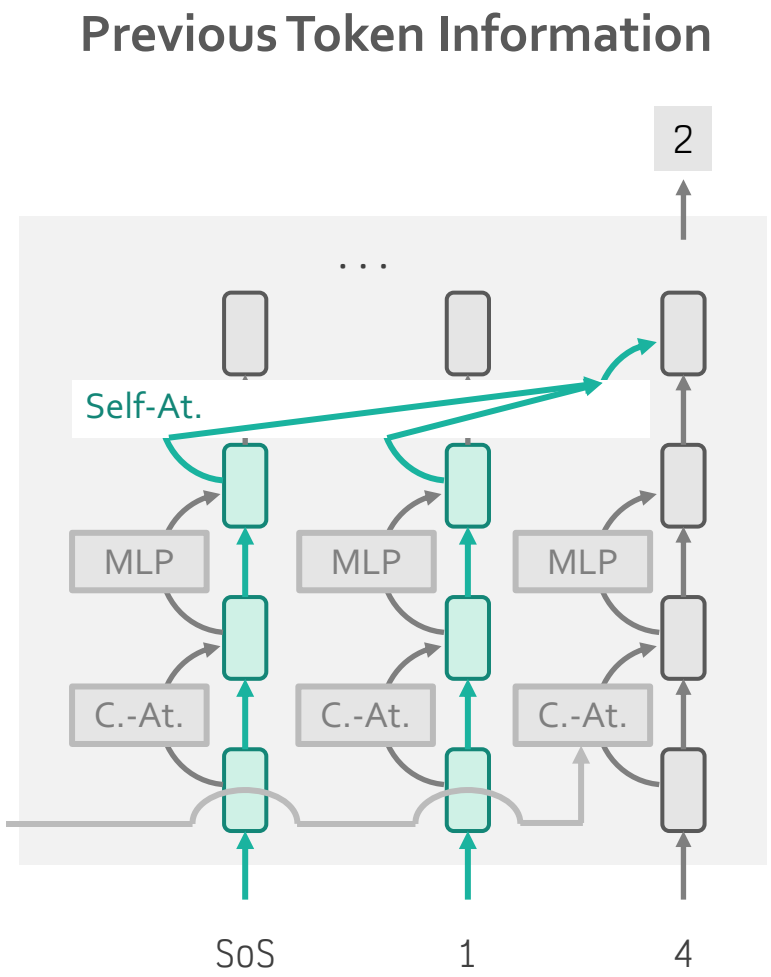


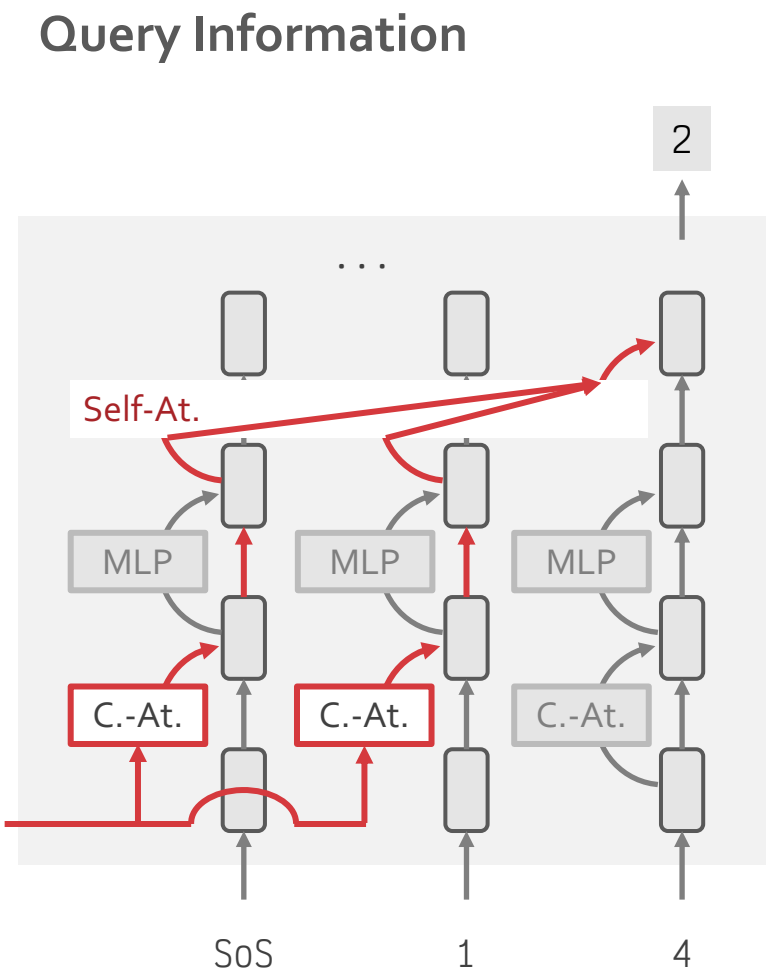
Focus only on the last prediction token

Information Movement through Self-Attention

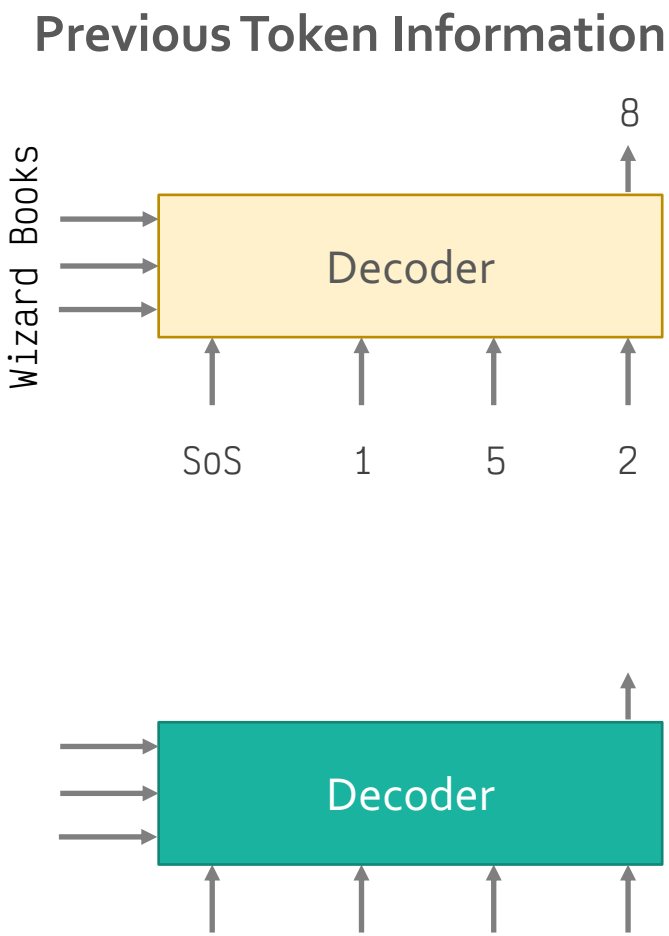


Information Movement through Self-Attention

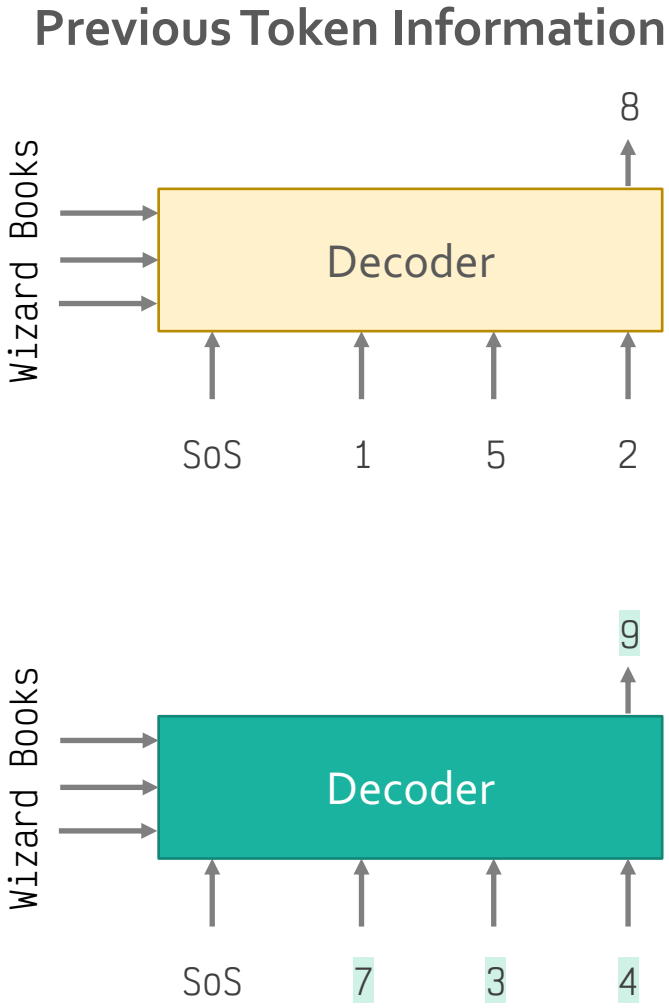




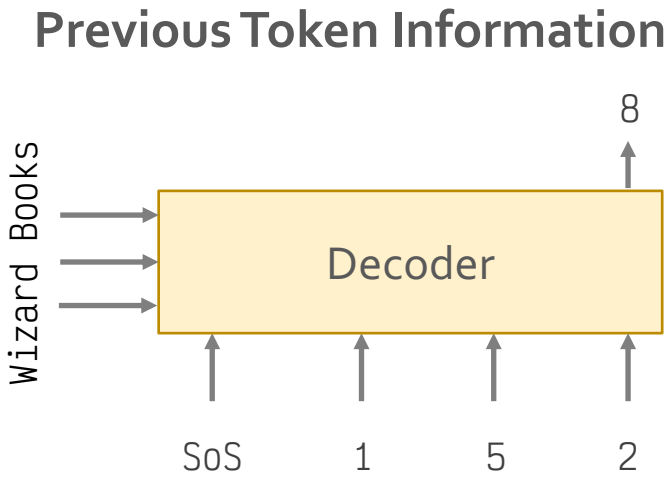
Two Patching Experiments

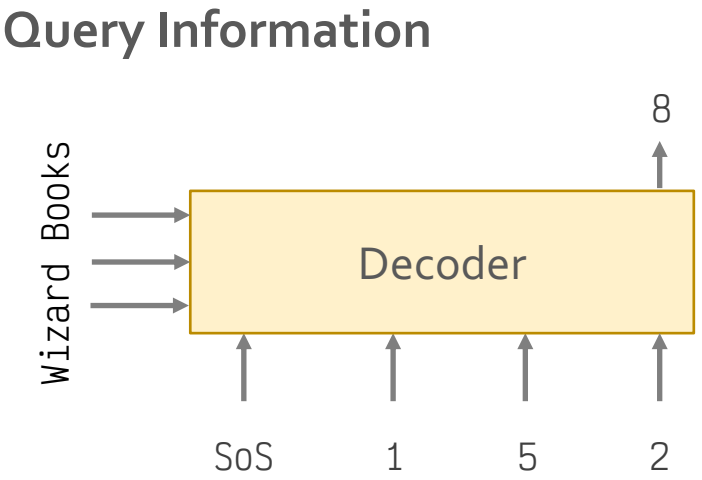


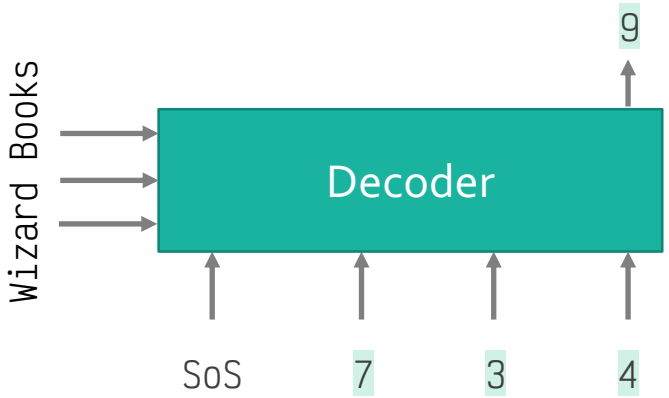
Two Patching Experiments

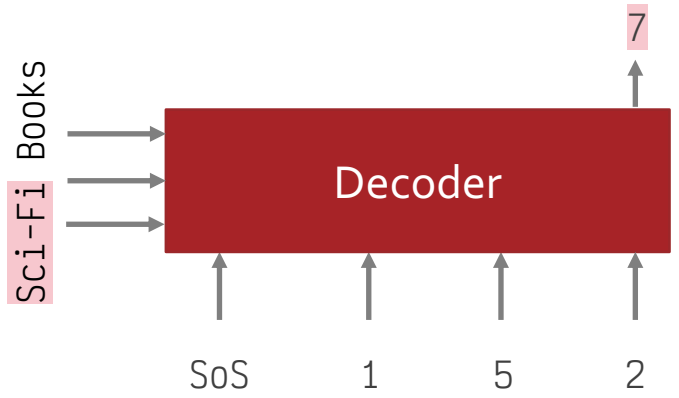


Two Patching Experiments



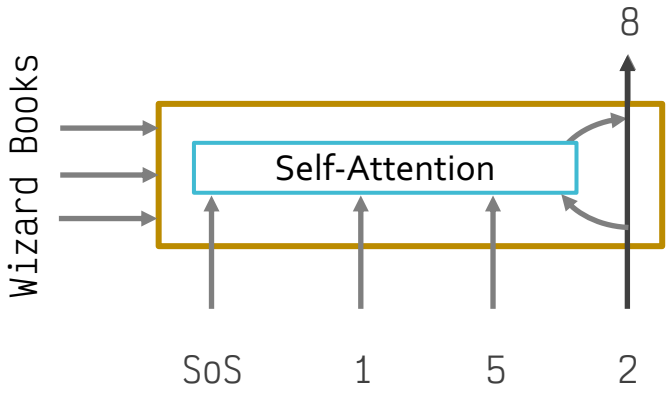


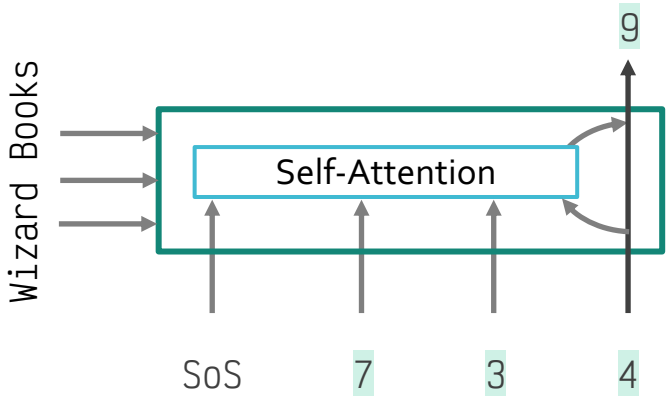




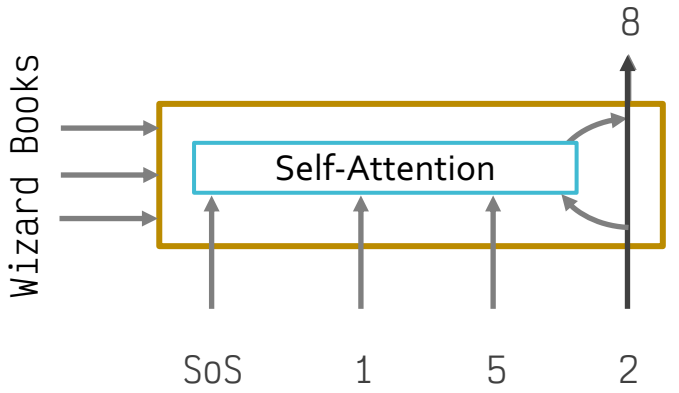
Two Patching Experiments

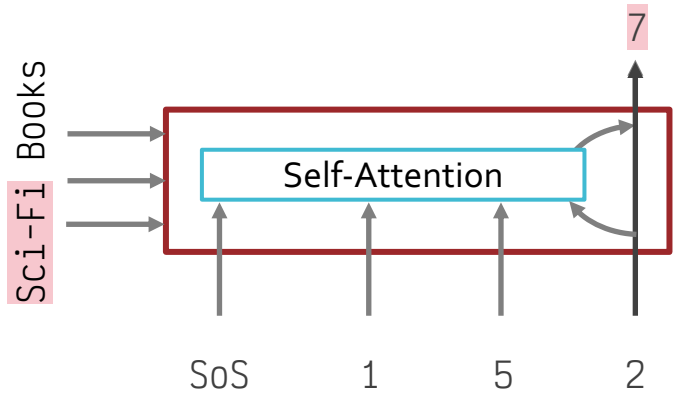
Previous Token Information



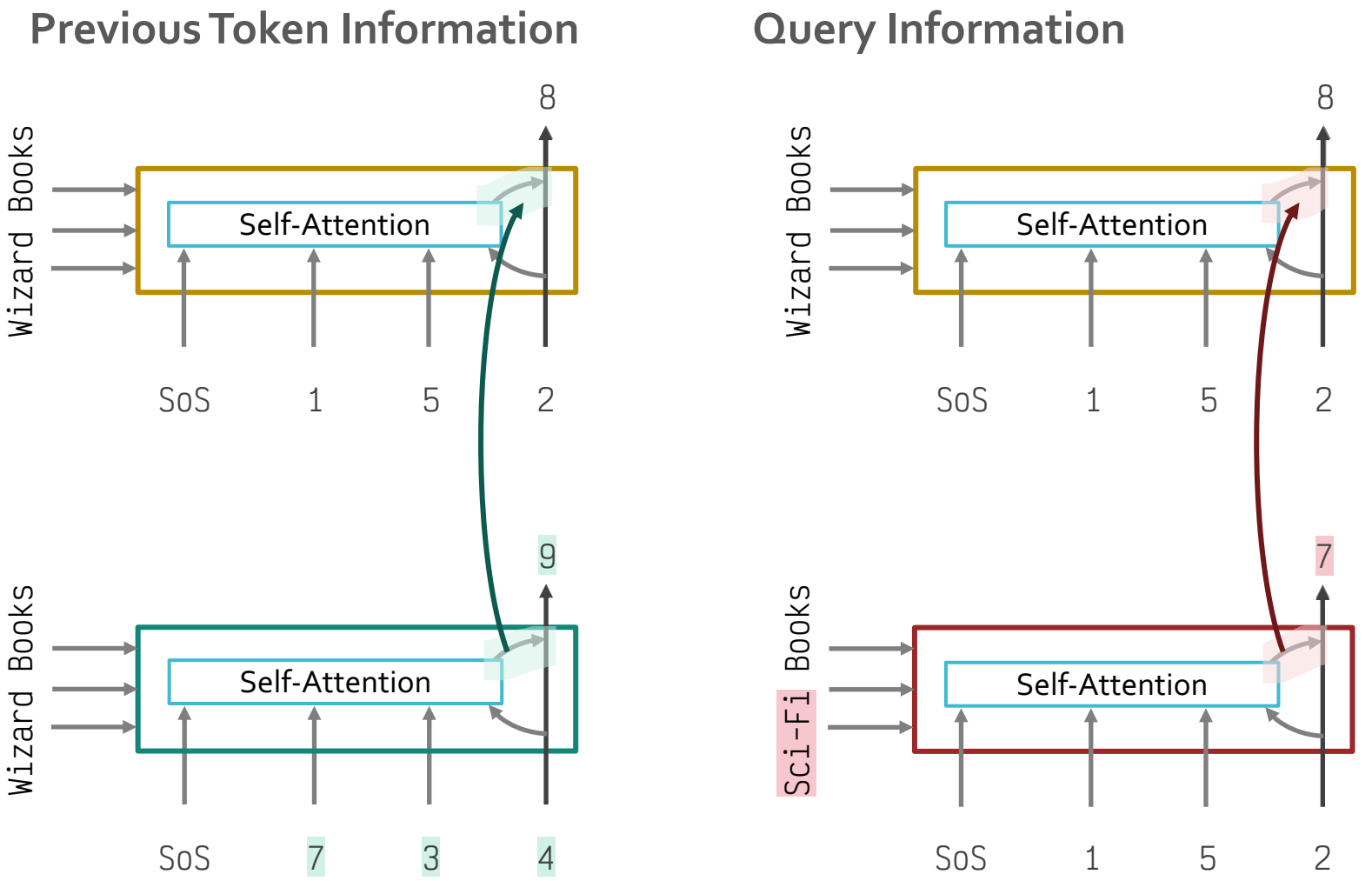


Query Information

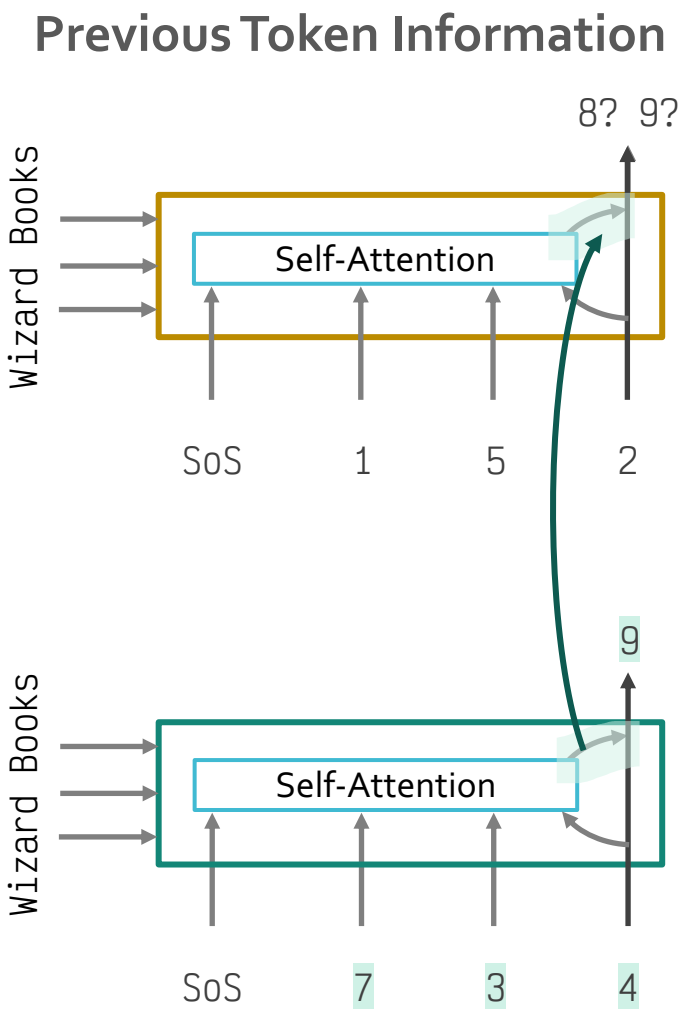


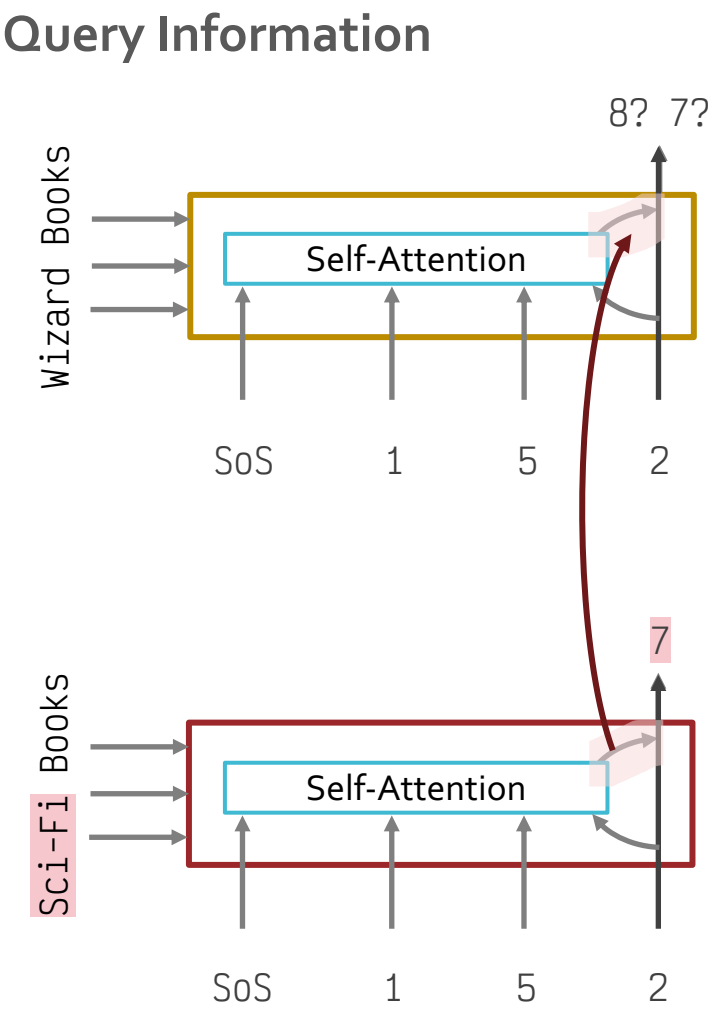


Two Patching Experiments

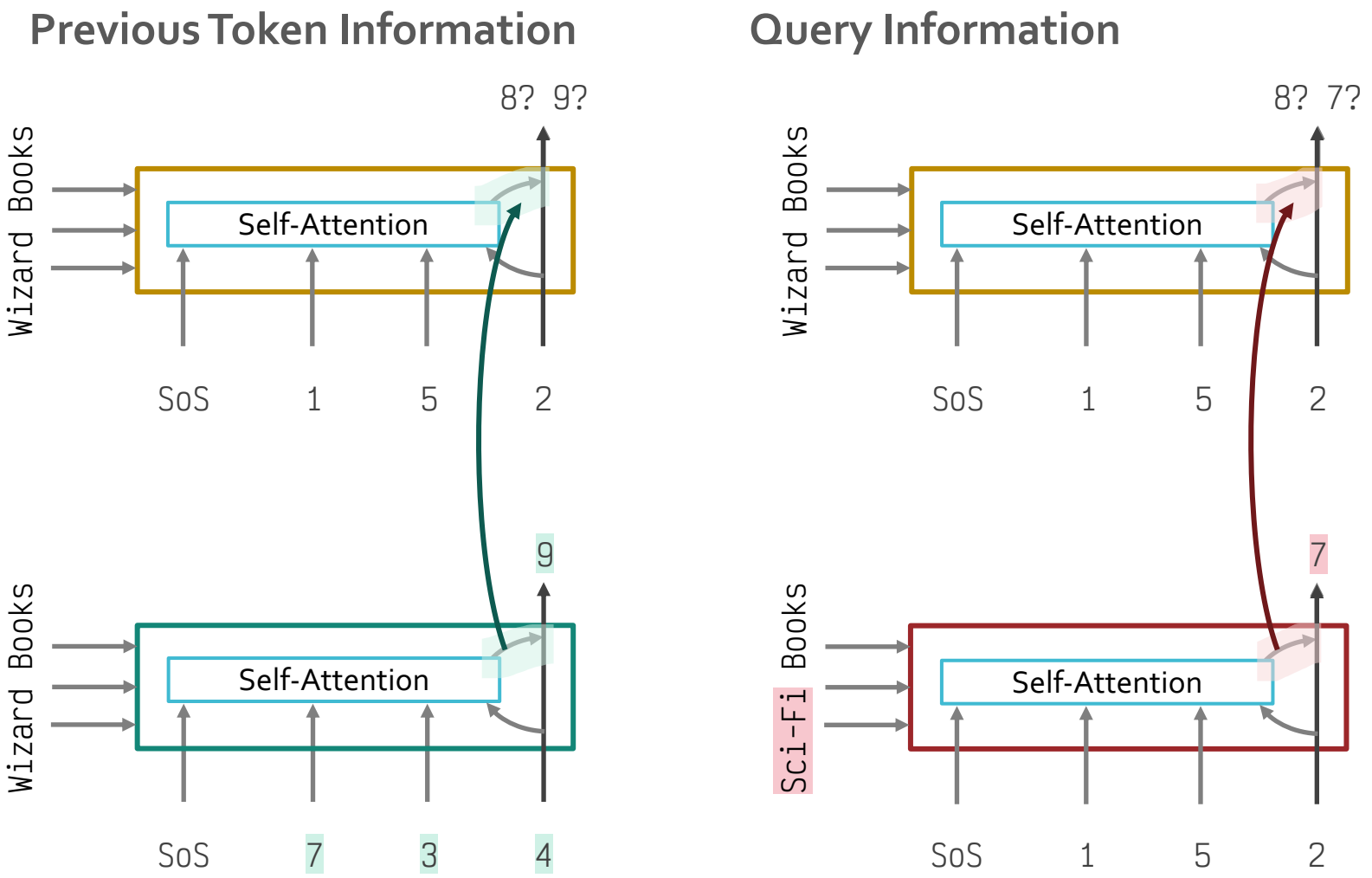


Two Patching Experiments



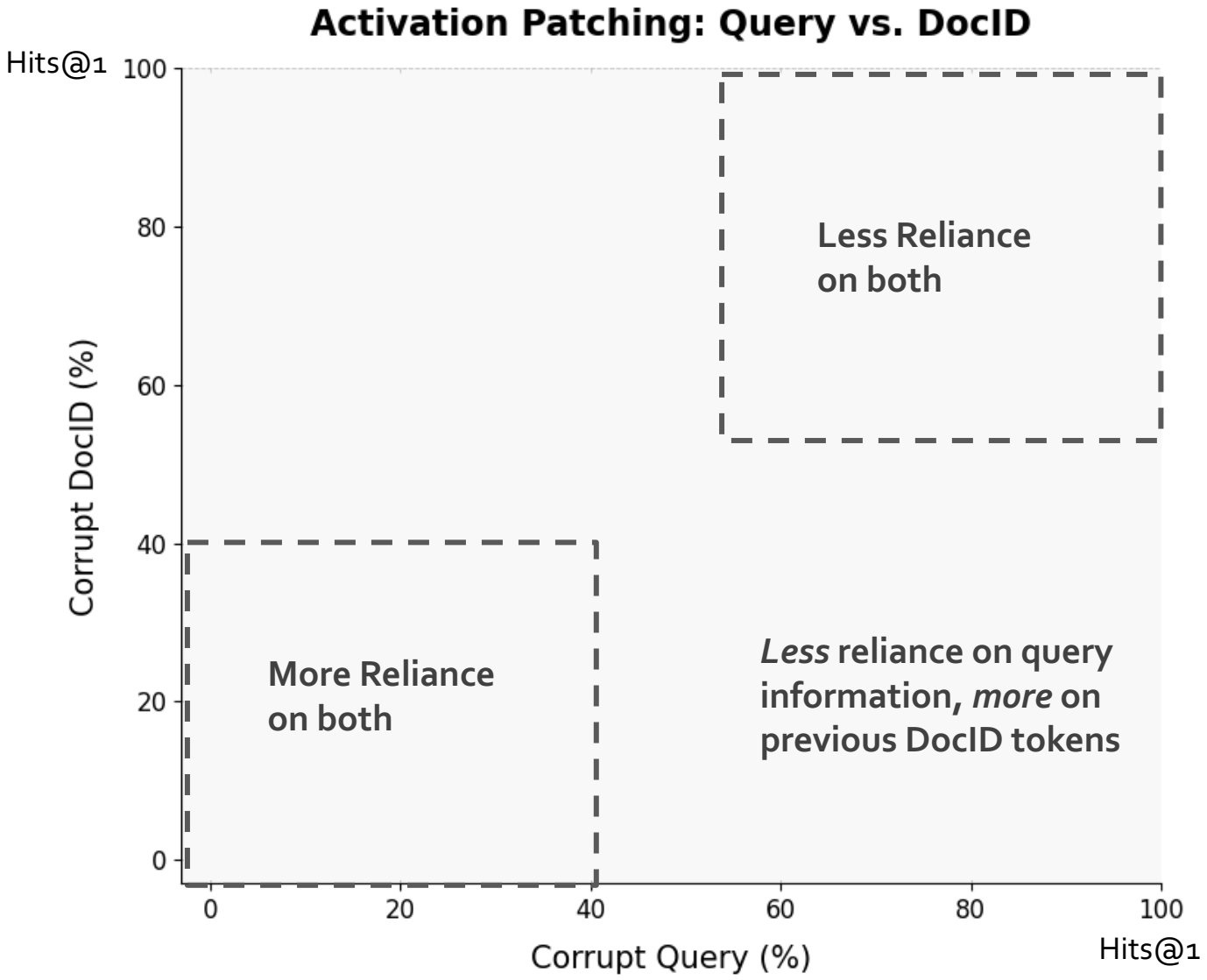


Two Patching Experiments



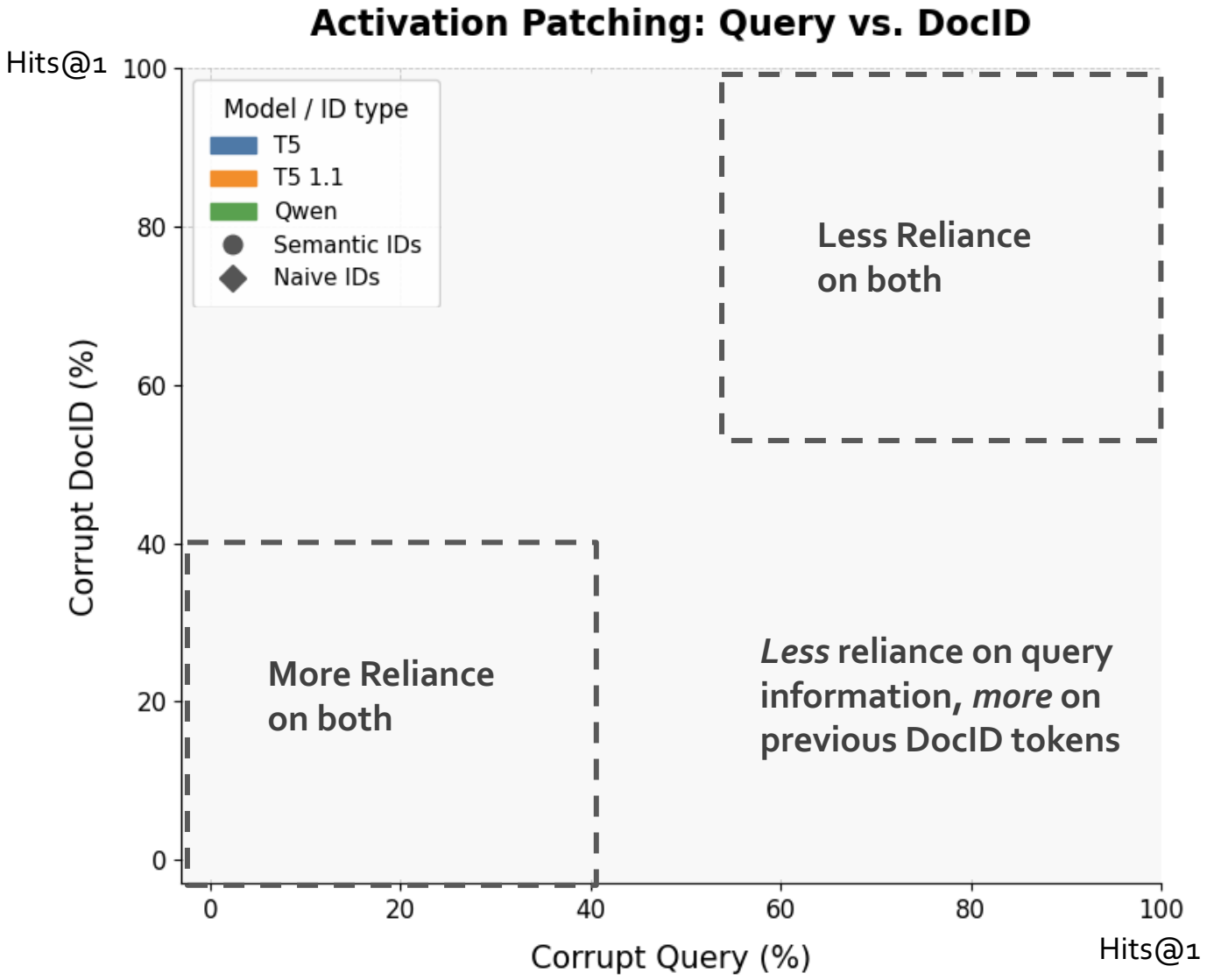
Worse Performance = More Reliance

Two Patching Experiments



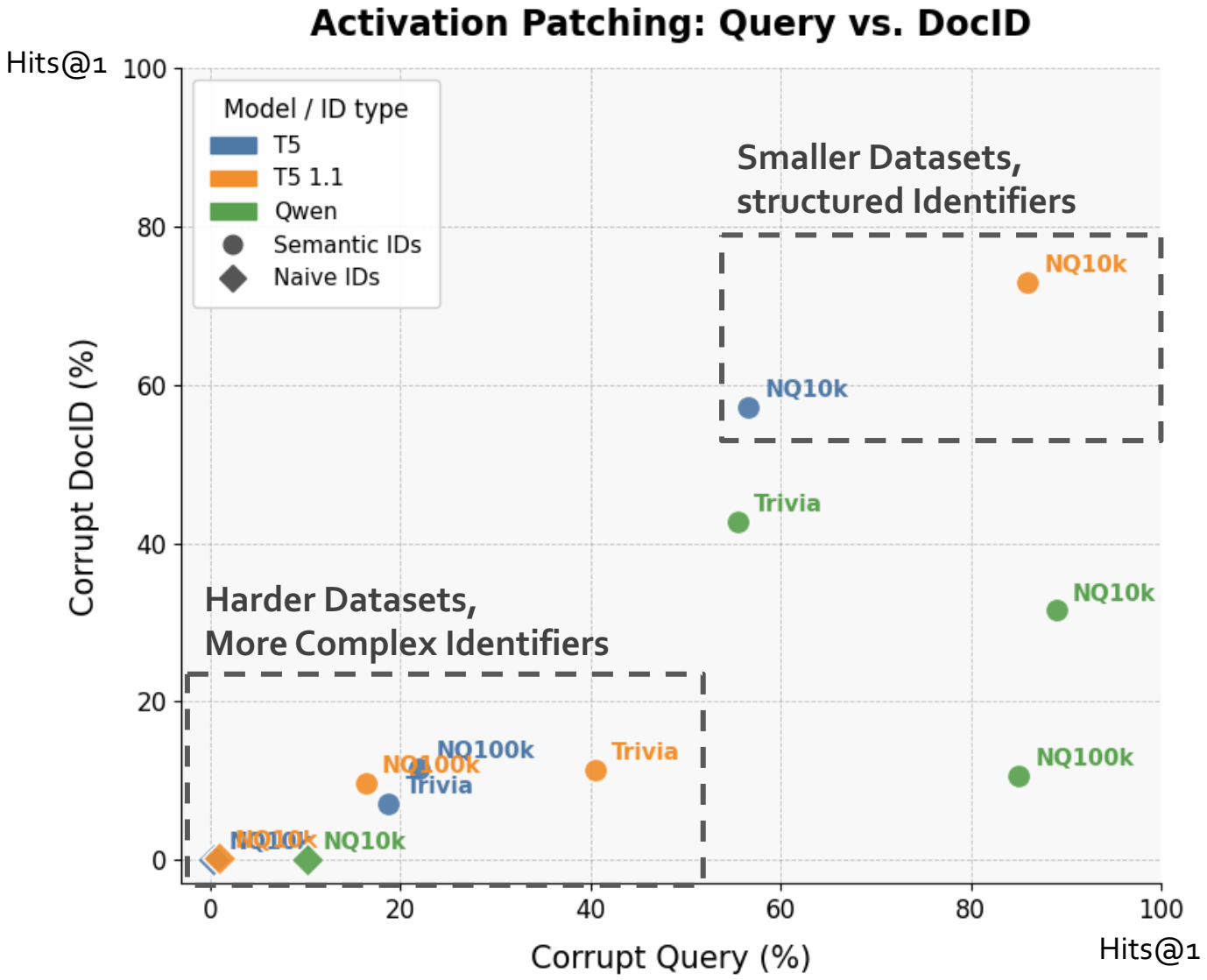
Worse Performance = More Reliance

Two Patching Experiments



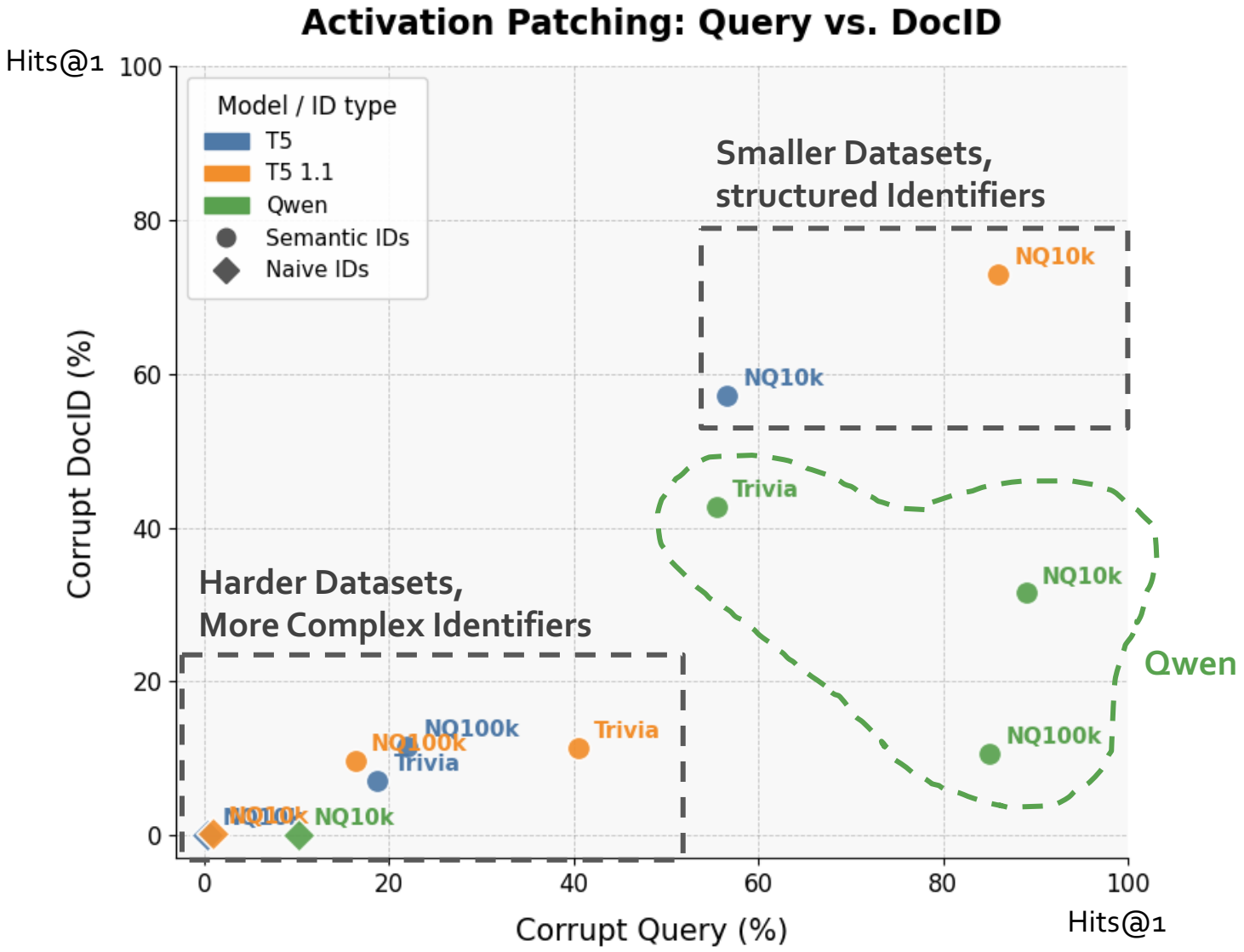
Worse Performance = More Reliance

Two Patching Experiments



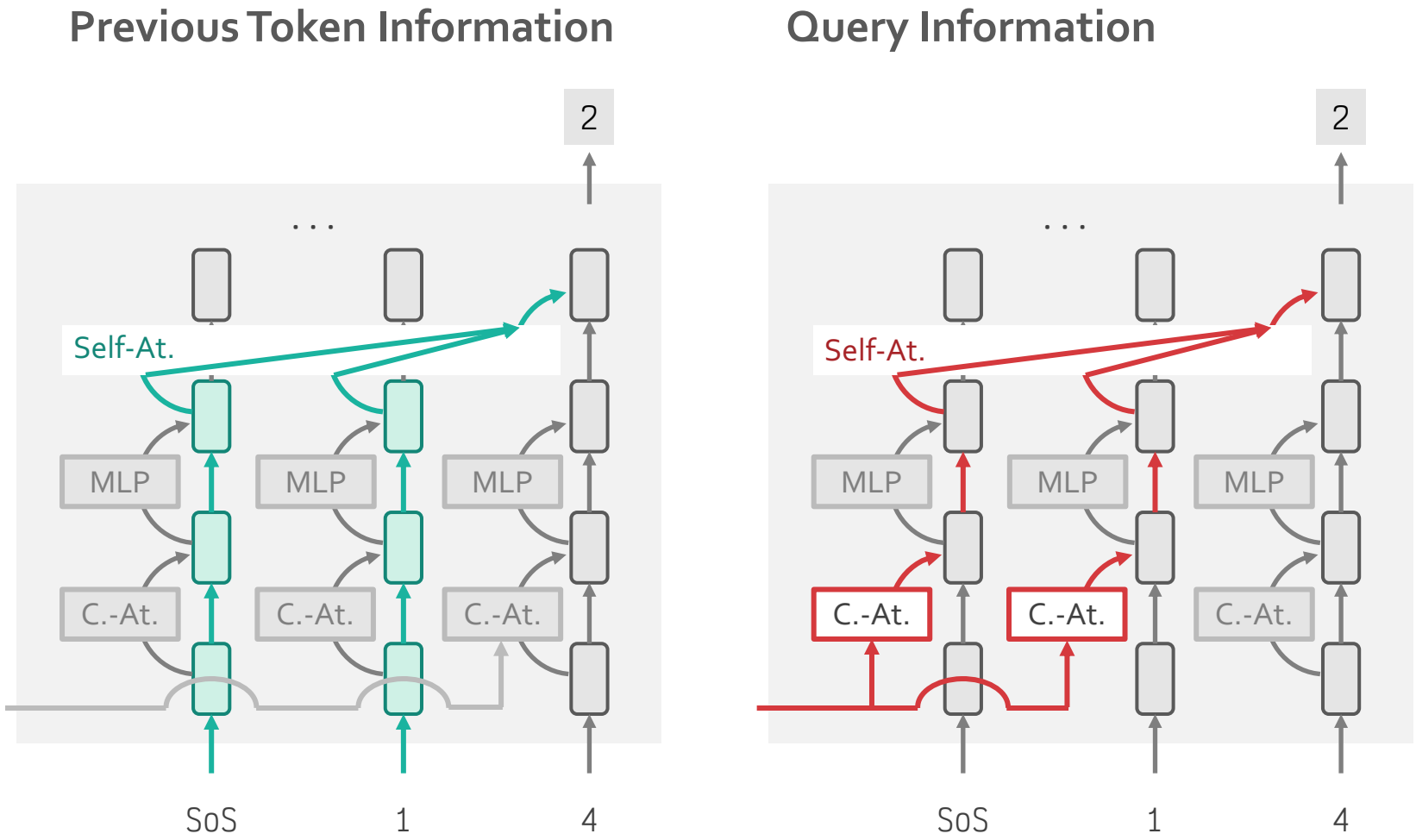
Worse Performance = More Reliance

Two Patching Experiments



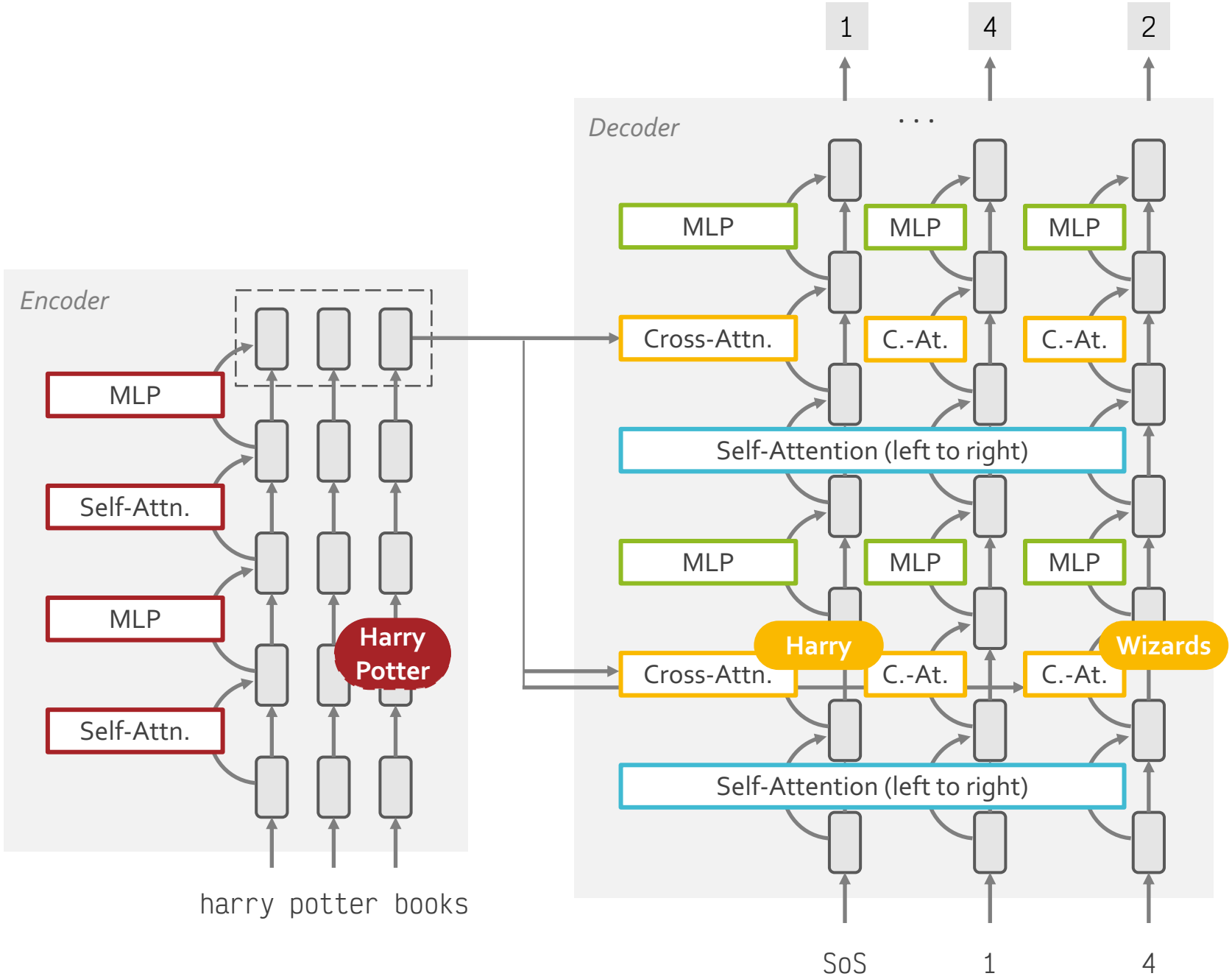
Worse Performance = More Reliance

Two Possibilities

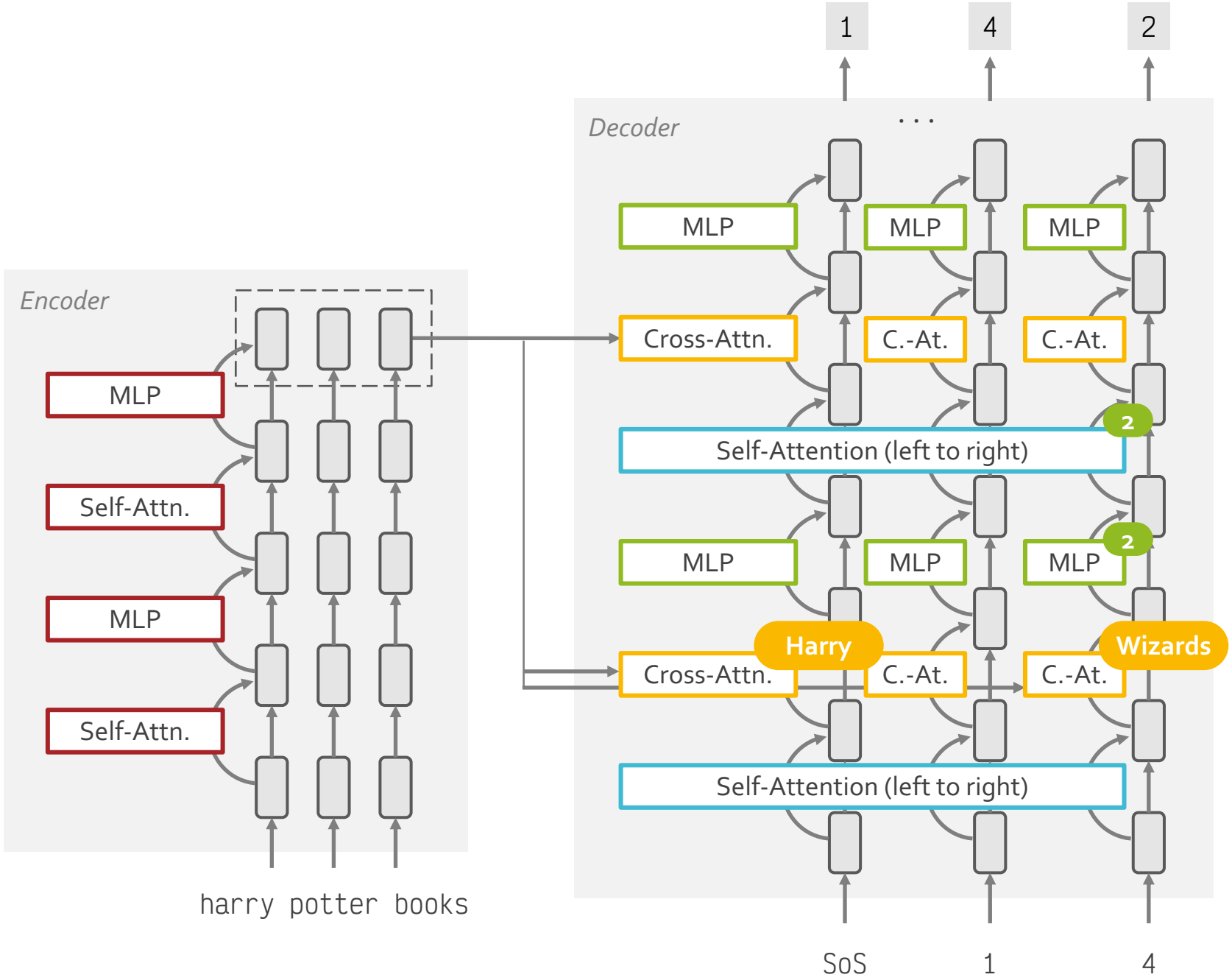


Self-Attention moves both information!

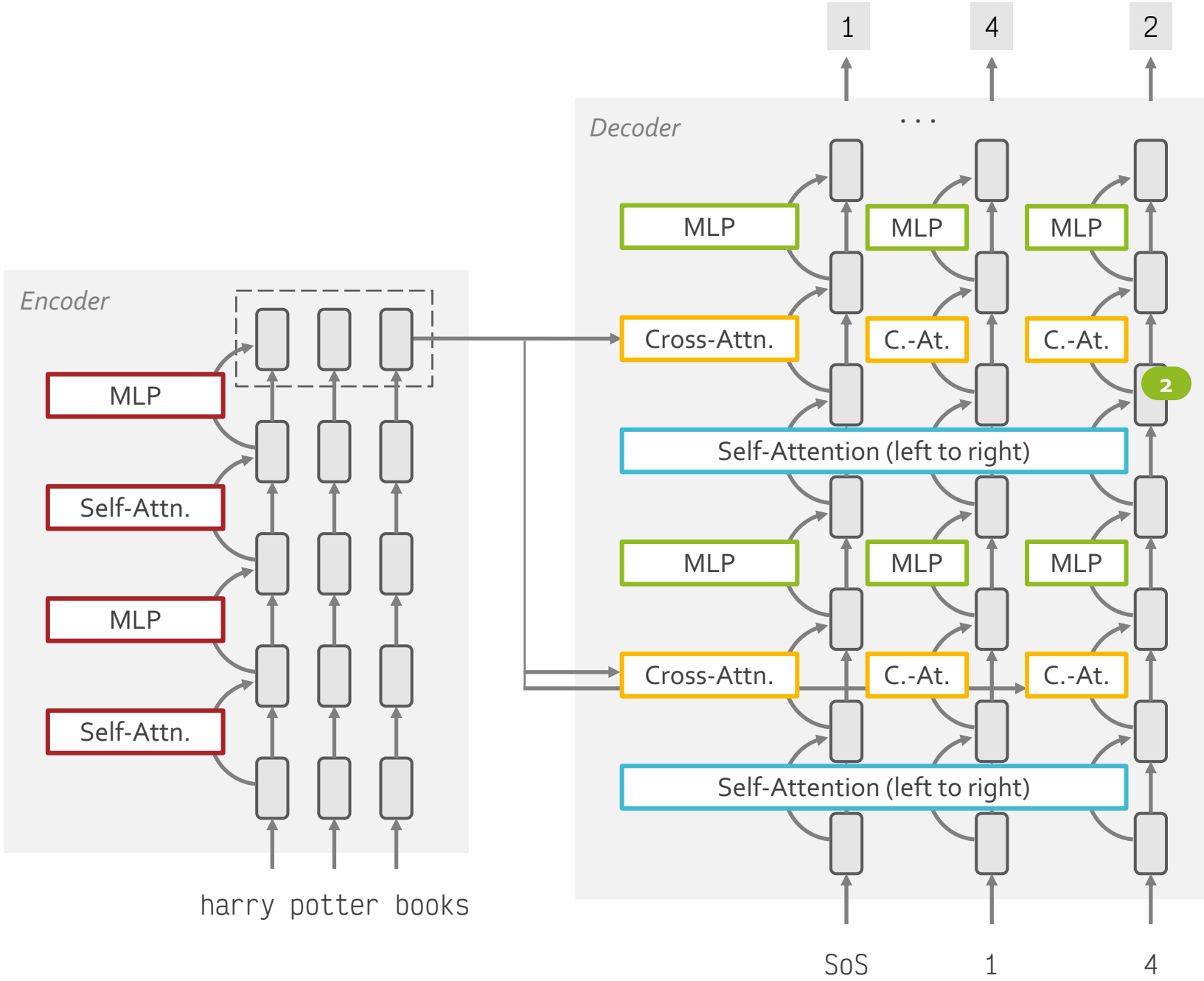
Full Picture



Full Picture

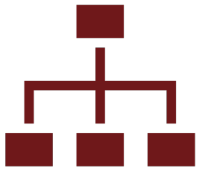


Full Picture



Summary

Analysis of Generative Retrieval Models



Adaption to Multi-Token Models

Faster, better and transparent retrieval

Interpretability of Generative Information Retrieval

Anja Reusch – Postdoctoral Researcher – NLP Lab, Technion

TECHNION