NEURAL INVERTED INDEX

RUCCI EMANUELE , GIANMARCO SCARANO

Deep Learning Course Held by Prof. Fabrizio Silvestri

SAPIENZA
UNIVERSITÀ DI ROMA

INDEX

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

INDEX

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

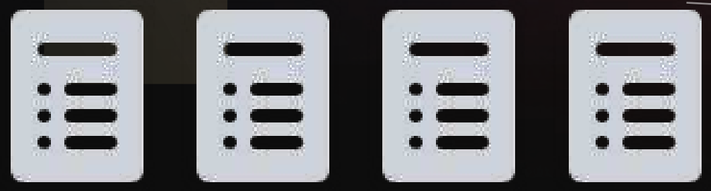
Information Retrieval and

DIFFERENTIABLE SEARCH INDEX

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

IR & DSI



Documents

Information

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

IR & DSI

Query

IR ENGINE

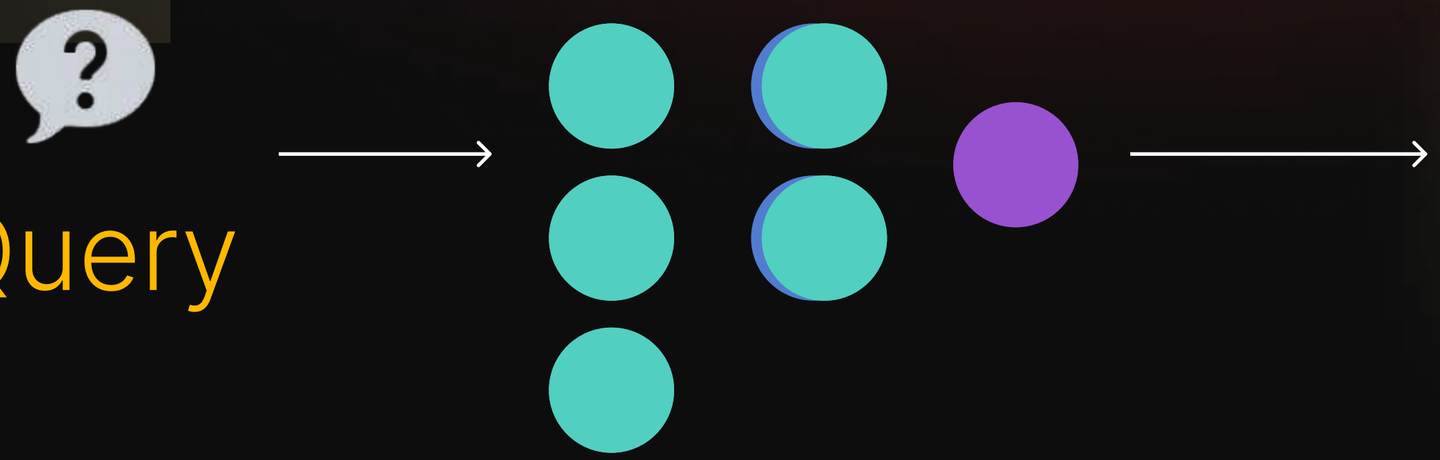
Information

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et ma...

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

IR & DSI



Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

IR & DSI

Query

Information

DOCUMENT IDENTIFIER

RETRIEVAL

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

IR & DSI

Information

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et ma...

INDEXING

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

What to Index?

DOCUMENT REPRESENTATION

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

DOCUMENT REP.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.
Aliquam auctor dolor et fermentum tincidunt. Nullam sagittis urna in urna aliquet consequat. Nunc laoreet porta nisl, et interdum elit vehicula at. Suspendisse maximus laoreet libero. Aliquam sit amet ultricies mauris. Nunc rhoncus elementum ex, ac sodales neque consequat sit amet. Donec dictum rhoncus urna non suscipit. Morbi ac orci vel urna rhoncus mollis eu quis mauris. Morbi ultrices blandit sodales. Quisque ut imperdiet orci. Nulla ornare non leo a porta. Morbi luctus, nulla ut eleifend suscipit, erat nisl ornare augue, vitae fermentum nunc tellus eu justo. Morbi dictum consequat mi et blandit. Donec semper est et cursus pellentesque. Nunc ut diam eget tellus varius facilisis. Etiam sed iaculis odio.
Nullam interdum turpis

Direct Indexing

Select first ***L*** tokens.

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

DOCUMENT REP.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.
Aliquam auctor dolor et fermentum tincidunt. Nullam sagittis urna in urna aliquet consequat. Nunc laoreet porta nisl, et interdum elit vehicula at. Suspendisse maximus laoreet libero. Aliquam sit amet ultricies mauris. Nunc rhoncus elementum ex, ac sodales neque consequat sit amet. Donec dictum rhoncus urna non suscipit. Morbi ac orci vel urna rhoncus mollis eu quis mauris. Morbi ultrices blandit sodales. Quisque ut imperdiet orci. Nulla ornare non leo a porta. Morbi luctus, nulla ut eleifend suscipit, erat nisl ornare augue, vitae fermentum nunc tellus eu justo. Morbi dictum consequat mi et blandit. Donec semper est et cursus pellentesque. Nunc ut diam eget tellus varius facilisis. Etiam sed iaculis odio.
Nullam interdum turpis

BUT, WHAT IF

this chunk misses the needed informations?

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

DOC BER

Lorem ipsum dolor s
dictum, lectus et im
aliquet velit justo sit
rutrum turpis varius
Vestibulum ullamcor
habitant morbi tristi
ac turpis egestas. P
ultrices imperdiet ar
pellentesque eros v
erat, et pretium dolo
nisl, non ultricies nu
dignissim ipsum nor
lacus volutpat, feug
Aliquam auctor dolo
urna in urna aliquet
interdum elit vehicu
Aliquam sit amet ult
ac sodales neque c
urna non suscipit. M
mauris. Morbi ultrice
orci. Nulla ornare nc
suscipit, erat nisl or
justo. Morbi dictum
et cursus pellentesc
Etiam sed iaculis od
Nullam interdum tur

ations?

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

DOCUMENT REP.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.
Aliquam auctor dolor et fermentum tincidunt. Nullam sagittis urna in urna aliquet consequat. Nunc laoreet porta nisl, et interdum elit vehicula at. Suspendisse maximus laoreet libero. Aliquam sit amet ultricies mauris. Nunc rhoncus elementum ex, ac sodales neque consequat sit amet. Donec dictum rhoncus urna non suscipit. Morbi ac orci vel urna rhoncus mollis eu quis mauris. Morbi ultrices blandit sodales. Quisque ut imperdiet orci. Nulla ornare non leo a porta. Morbi luctus, nulla ut eleifend suscipit, erat nisl ornare augue, vitae fermentum nunc tellus eu justo. Morbi dictum consequat mi et blandit. Donec semper est et cursus pellentesque. Nunc ut diam eget tellus varius facilisis. Etiam sed iaculis odio.
Nullam interdum turpis

SET INDEXING

Remove:

- Stopwords
- Duplicates

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

DOC

NG

Lorem ipsum dolor sit amet
dictum, lectus et imperdiet
aliquet velit justo sit amet l
rutrum turpis varius quis. S
Vestibulum ullamcorper lao
habitant morbi tristique ser
ac turpis egestas. Phasellus
ultrices imperdiet arcu. Mau
pellentesque eros vulputate
erat, et pretium dolor males
nisl, non ultricies nunc blan
dignissim ipsum non euism
lacus volutpat, feugiat ullam
Aliquam auctor dolor et fer
urna in urna aliquet conseq
interdum elit vehicula at. Su
Aliquam sit amet ultricies m
ac sodales neque consequa
urna non suscipit. Morbi ac
mauris. Morbi ultrices blanc
orci. Nulla ornare non leo a
suscipit, erat nisl ornare au
justo. Morbi dictum conseq
et cursus pellentesque. Nu
Etiam sed iaculis odio.
Nullam interdum turpis

NINJAS

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

DOCUMENT REP.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.
Aliquam auctor dolor et fermentum tincidunt. Nullam sagittis urna in urna aliquet consequat. Nunc laoreet porta nisl, et interdum elit vehicula at. Suspendisse maximus laoreet libero. Aliquam sit amet ultricies mauris. Nunc rhoncus elementum ex, ac sodales neque consequat sit amet. Donec dictum rhoncus urna non suscipit. Morbi ac orci vel urna rhoncus mollis eu quis mauris. Morbi ultrices blandit sodales. Quisque ut imperdiet orci. Nulla ornare non leo a porta. Morbi luctus, nulla ut eleifend suscipit, erat nisl ornare augue, vitae fermentum nunc tellus eu justo. Morbi dictum consequat mi et blandit. Donec semper est et cursus pellentesque. Nunc ut diam eget tellus varius facilisis. Etiam sed iaculis odio.
Nullam interdum turpis

SUMMARIZATION

Falcon Summarization

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

How to Index

DOC ID REPRESENTATION

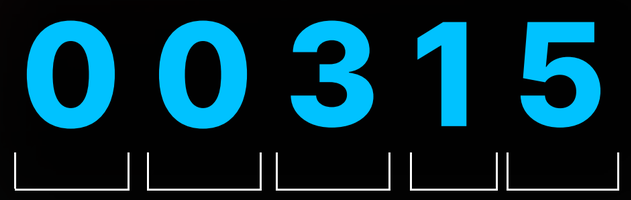
Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

DocID REP.

Naively Structured

In a **Generative setting** the DocID
is treated as a tokenizable string.



Unstructured Atomic

In a **Discriminative setting** the
DocID has to be chosen directly.

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

DocID REP.

Conceptual DocID

Treat a DocID as a tokenizable hierarchical string using external knowledge-base, such as WikiData or BabelNet, like:

Prefix + coarse_grained_ID + fine_grained_ID + doc_chunk_ID

docID_fruit_apple_03

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

Training Type

DISCRIMINATIVE

GENERATIVE

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

TRAINING TYPE

DISCRIMINATIVE

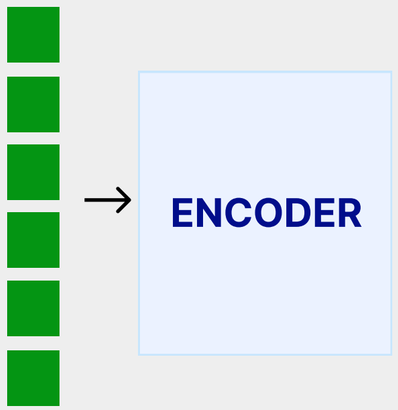
TOKENS

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

TRAINING TYPE

DISCRIMINATIVE

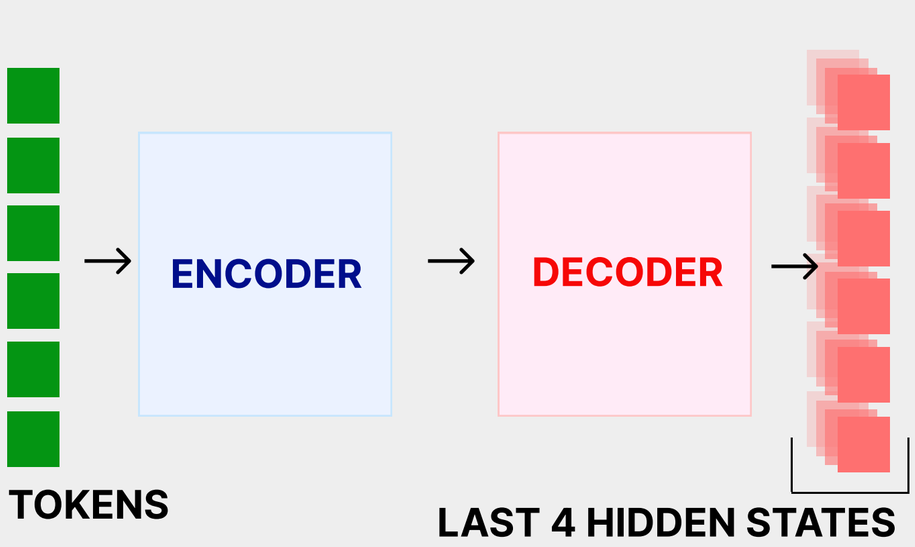


Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

TRAINING TYPE

DISCRIMINATIVE

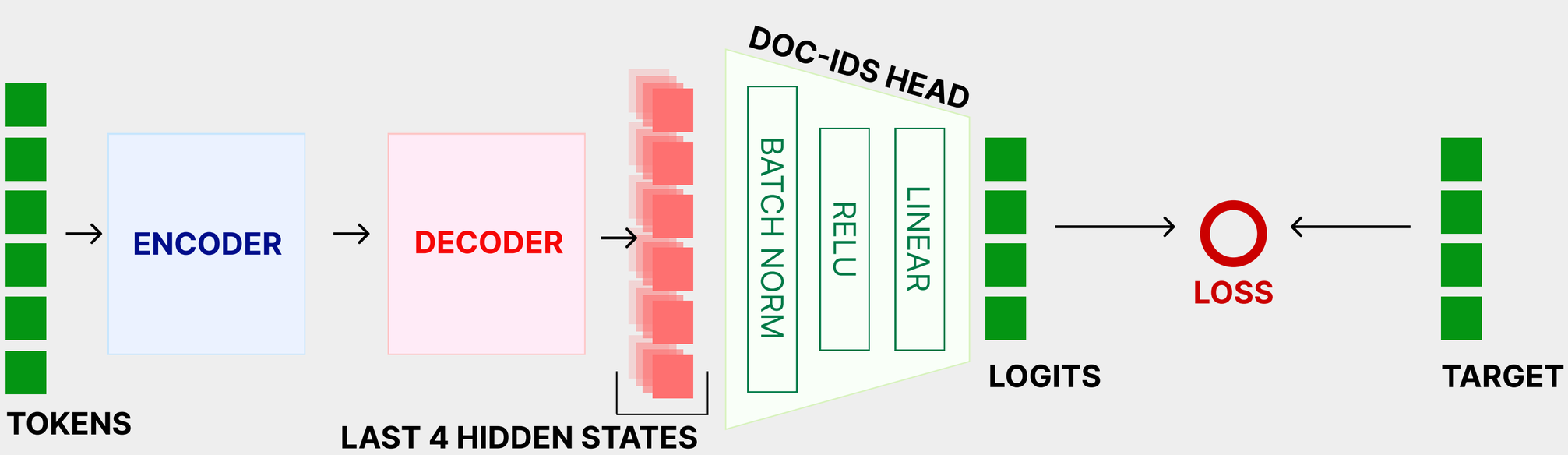


Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

TRAINING TYPE

DISCRIMINATIVE

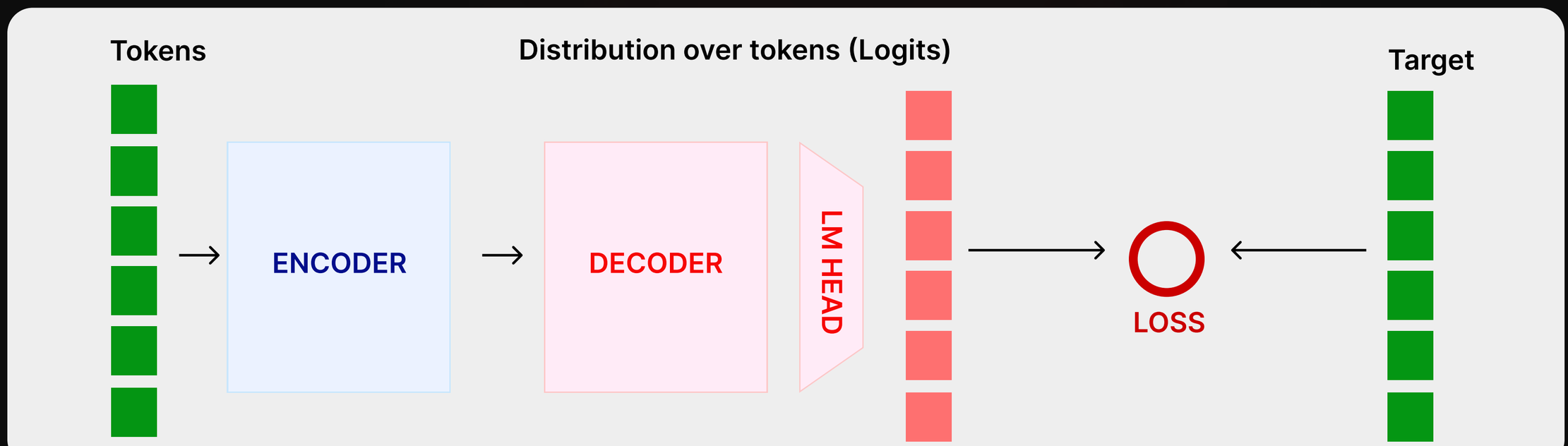


Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

TRAINING TYPE

GENERATIVE

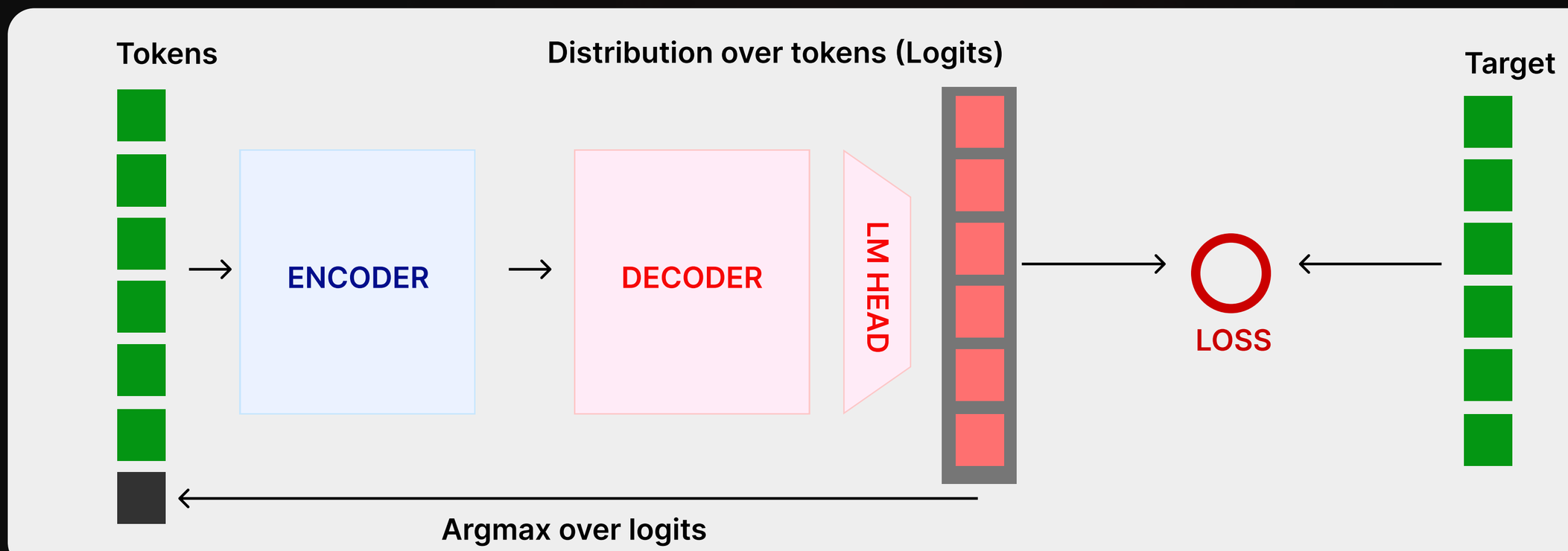


Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

TRAINING TYPE

GENERATIVE



Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

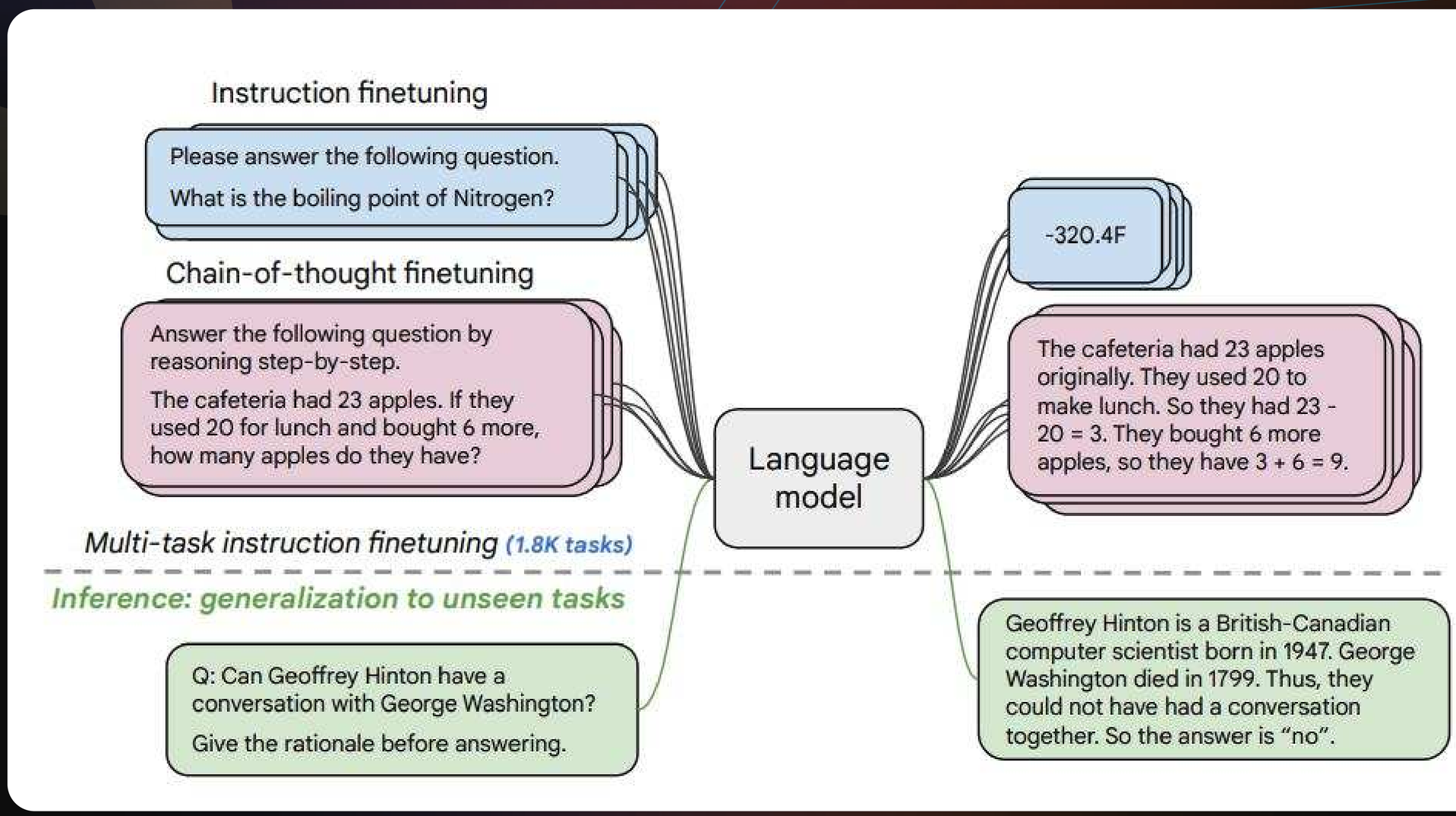
Architecture Design

FLAN-T5, BERT & MoE

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

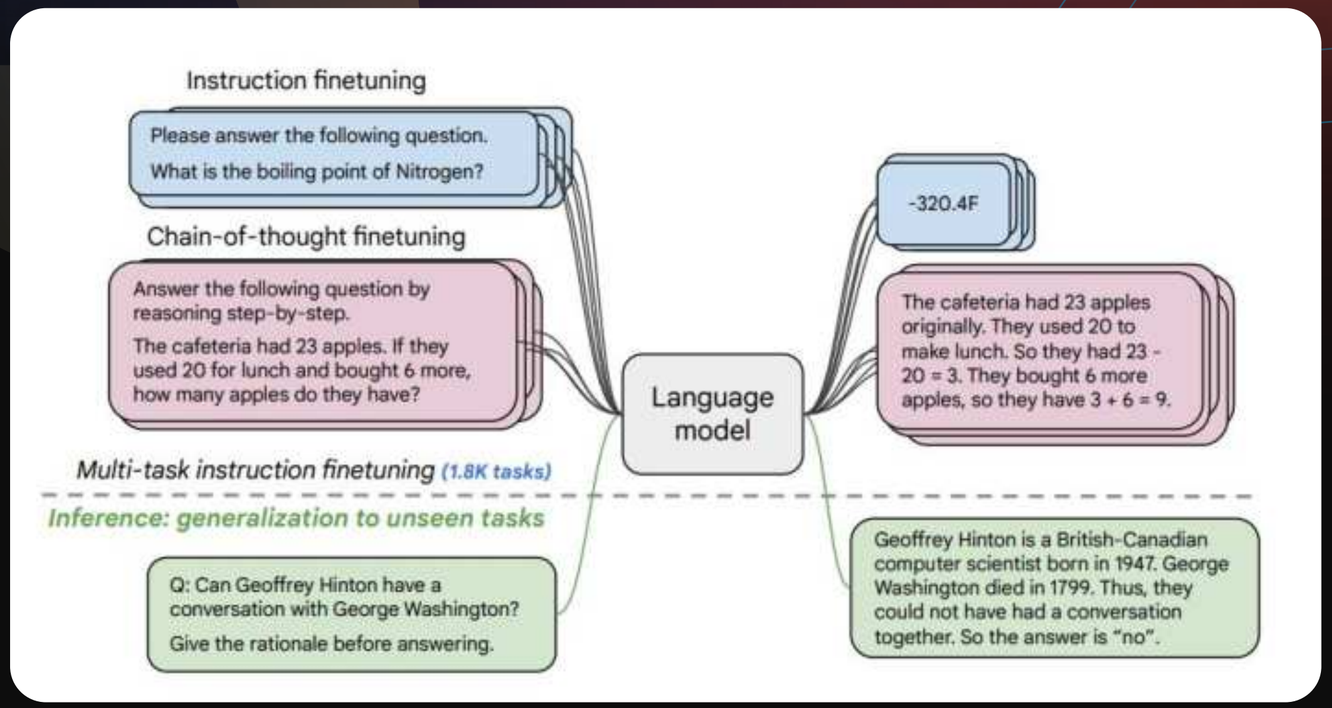
BASELINE: Flan-T5



Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

BASELINE: Flan-T5



1800 Tasks

248M Params

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

BERT: Shared parameters

Shared parameters leads to more efficient computations?

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

BERT: Shared parameters

Shared parameters leads to more efficient computations?

Yes, along with benefits of fewer parameters.

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

BERT: Shared parameters

More reasons?

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

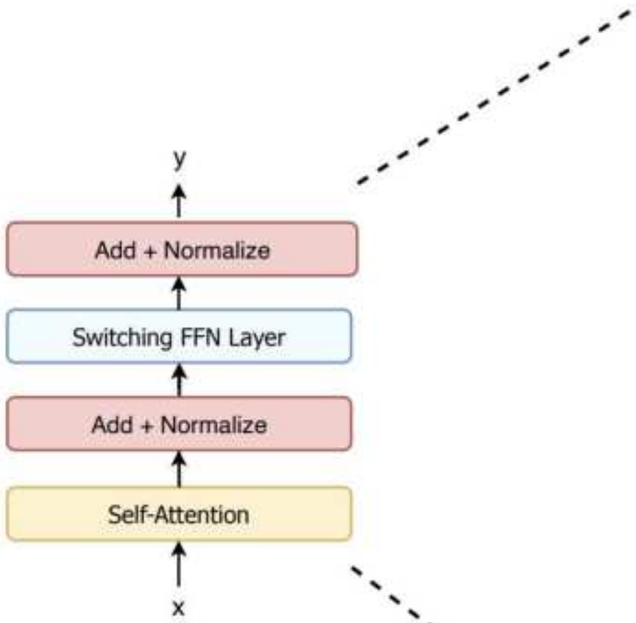
BERT: Shared parameters

What is the impact of parameters sharing on Index and Retrieval?

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

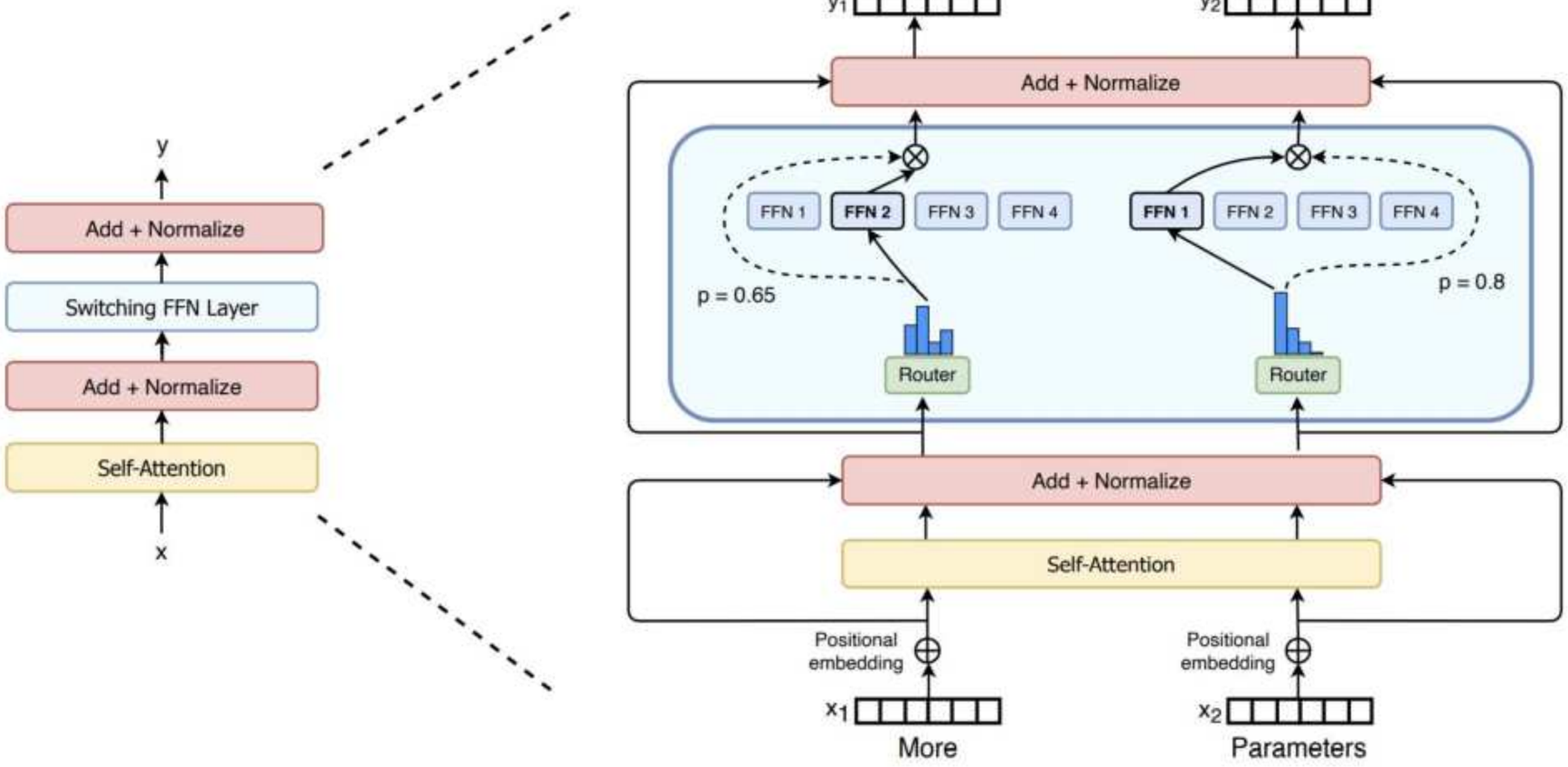
MoE + Flan-T5



Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

MoE + Flan-T5



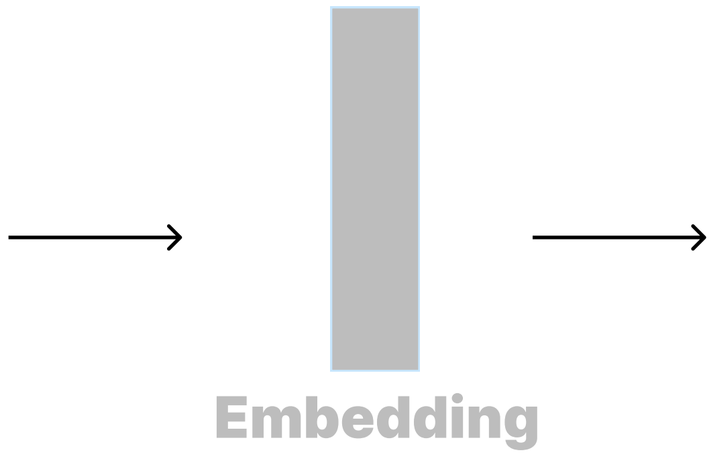
Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

MoE + Flan-T5

Switch Transformer

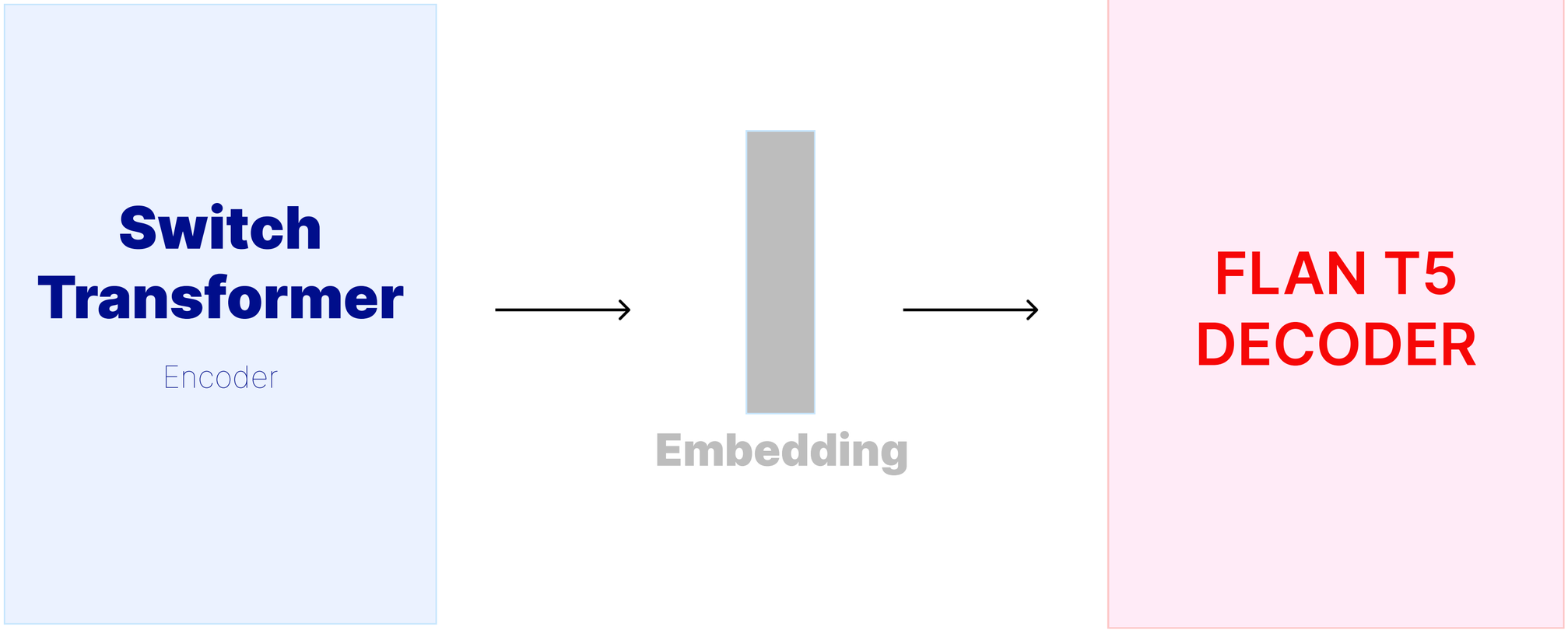
Encoder



Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

MoE + Flan-T5



Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

MoE + Flan-T5

Does the MoE specialize on certain aspects of the documents?

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

Technical Tricks

QUANTIZATION & FINETUNING

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

QUANTIZATION

Before Quantization

FLAN-T5

248 M

Parameters

→ ⟶

BERT

139 M

Parameters

MoE

868 M

Parameters

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

QUANTIZATION

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

LoRA

$$W_0 + \Delta W = W_0 + BA$$

Large Model

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

LoRA

$$W_0 + \Delta W = W_0 + BA$$

Large Model

$$B \in \mathbb{R}^{dxr}$$

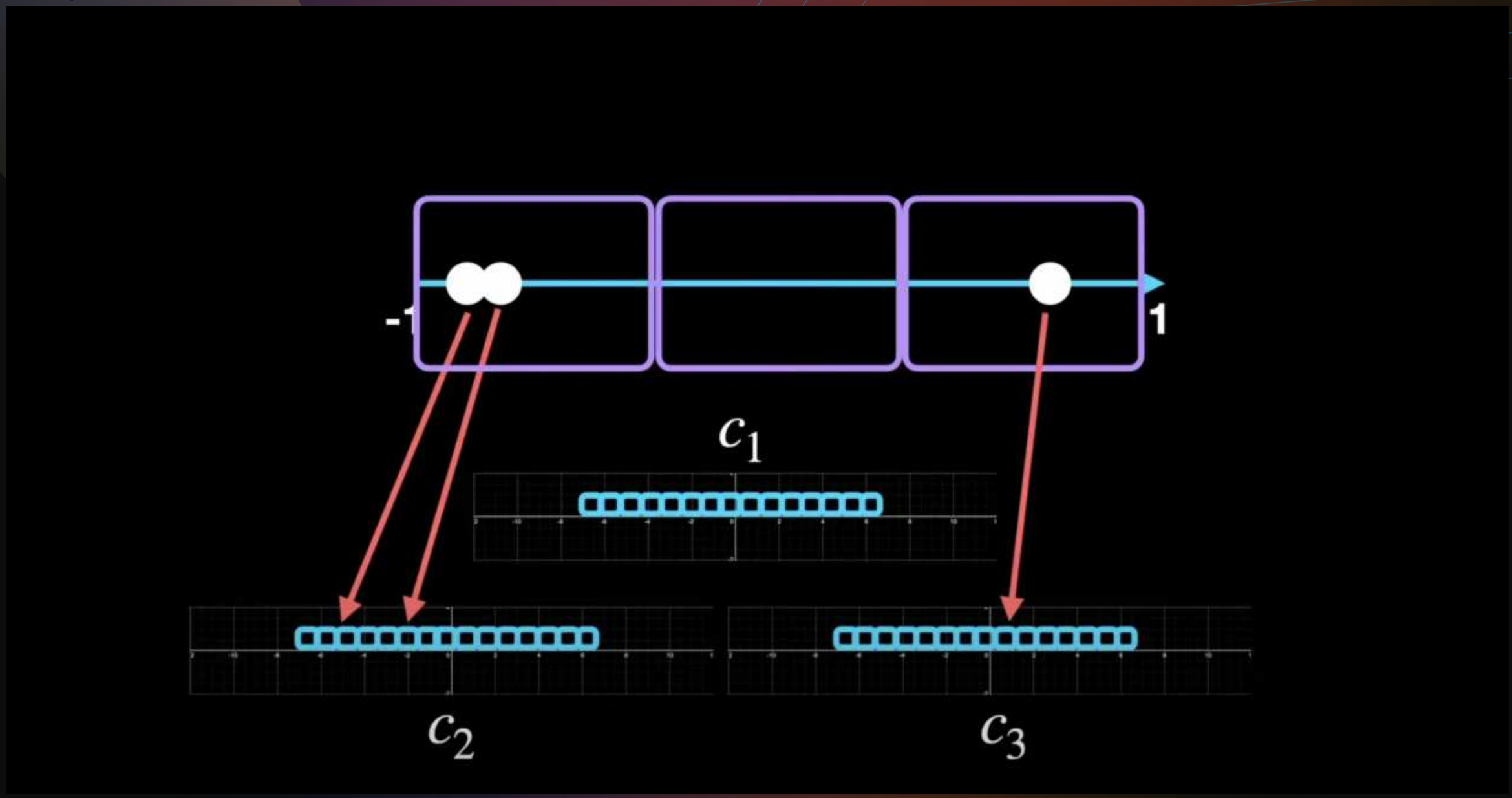
$$A \in \mathbb{R}^{rxk}$$

$$rank\ r \ll \min(d, k)$$

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

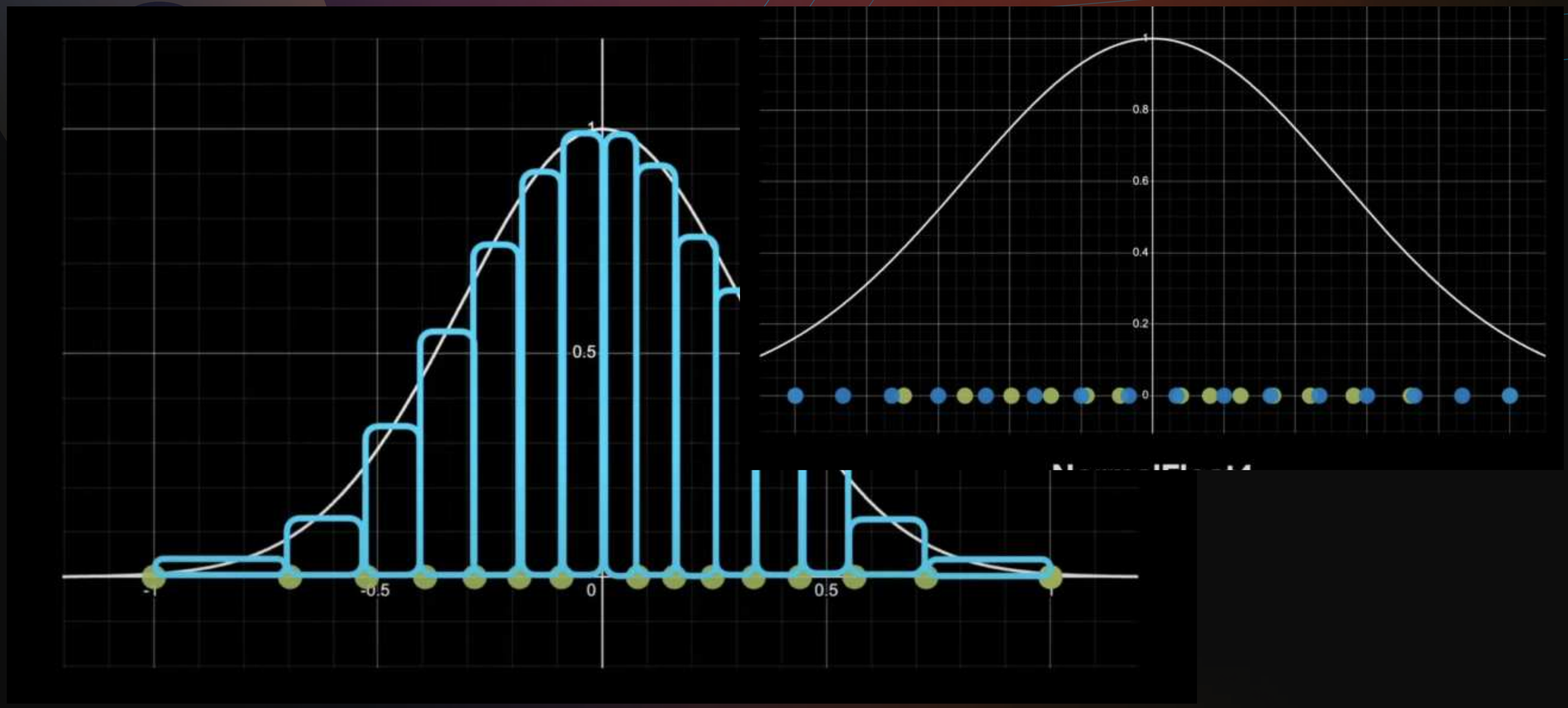
Q-LoRA: NF4



Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

Q-LoRA: NF4



Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

QUANTIZATION

After Quantization

FLAN-T5

1.3 M

Parameters

→ →

BERT

1.3 M

Parameters

MoE

2.6 M

Parameters

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

FINETUNING

FLAN-T5

66 M
Trainable

182 M
Non trainable

Update

4
Last layers of encoder and decoder

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

Multitask Setup

INSTRUCTION & RATIO TASKS

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

INSTRUCTION

Prefix + Input

INDEXING →

Generate a document identifier with 5 digits between 0 and 9 for the following document:

RETRIEVAL

Generate a document identifier with 5 digits between 0 and 9 as response to the following query:

Emanuele Rucci, Gianmarco Scarano

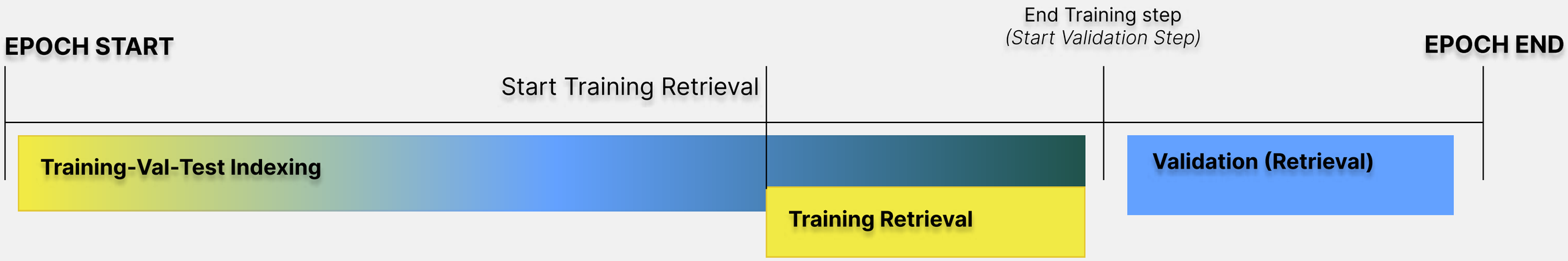
SAPIENZA
UNIVERSITÀ DI ROMA

TASKS RATIO

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

TASKS RATIO



Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

Dataset and Tests

Data organization & Experiments

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

DATASET'S INFORMATION

MS MARCO

Microsoft

General Info

3.21 M
Documents

59
Chunks

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

DATASET'S INFORMATION

We focused on

'00'

Chunk

200K

Documents

7773

Words mean per Doc

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

EXPERIMENTS

Table 1: Flan-T5, BERT and Encoder MoE-Decoder tests

Model	Doc Strategy	Doc ID Strategy	Prompting	Training Type	N. Docs	Fine-Tuning
Flan-T5	Direct Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Generative	50k	False
Flan-T5	Direct Indexing	Naively Structured	False	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Discriminative	10k	False
Flan-T5	Set Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	True
Flan-T5	Summarization	Naively Structured	True	Generative	200k	True
BERT	Direct Indexing	Naively Structured	True	Generative	10k	False
BERT	Summarization	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	False	Generative	10k	False

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

EXPERIMENTS

Table 1: Flan-T5, BERT and Encoder MoE-Decoder tests

Model	Doc Strategy	Doc ID Strategy	Prompting	Training Type	N. Docs	Fine-Tuning
Flan-T5	Direct Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Generative	50k	False
Flan-T5	Direct Indexing	Naively Structured	False	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Discriminative	10k	False
Flan-T5	Set Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	True
Flan-T5	Summarization	Naively Structured	True	Generative	200k	True
BERT	Direct Indexing	Naively Structured	True	Generative	10k	False
BERT	Summarization	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	False	Generative	10k	False

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

EXPERIMENTS

Table 1: Flan-T5, BERT and Encoder MoE-Decoder tests

Model	Doc Strategy	Doc ID Strategy	Prompting	Training Type	N. Docs	Fine-Tuning
Flan-T5	Direct Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Generative	50k	False
Flan-T5	Direct Indexing	Naively Structured	False	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Discriminative	10k	False
Flan-T5	Set Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	True
Flan-T5	Summarization	Naively Structured	True	Generative	200k	True
BERT	Direct Indexing	Naively Structured	True	Generative	10k	False
BERT	Summarization	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	False	Generative	10k	False

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

EXPERIMENTS

Table 1: Flan-T5, BERT and Encoder MoE-Decoder tests

Model	Doc Strategy	Doc ID Strategy	Prompting	Training Type	N. Docs	Fine-Tuning
Flan-T5	Direct Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Generative	50k	False
Flan-T5	Direct Indexing	Naively Structured	False	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Discriminative	10k	False
Flan-T5	Set Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	True
Flan-T5	Summarization	Naively Structured	True	Generative	200k	True
BERT	Direct Indexing	Naively Structured	True	Generative	10k	False
BERT	Summarization	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	False	Generative	10k	False

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

EXPERIMENTS

Table 1: Flan-T5, BERT and Encoder MoE-Decoder tests

Model	Doc Strategy	Doc ID Strategy	Prompting	Training Type	N. Docs	Fine-Tuning
Flan-T5	Direct Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Generative	50k	False
Flan-T5	Direct Indexing	Naively Structured	False	Generative	10k	False
Flan-T5	Direct Indexing	Naively Structured	True	Discriminative	10k	False
Flan-T5	Set Indexing	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	False
Flan-T5	Summarization	Naively Structured	True	Generative	10k	True
Flan-T5	Summarization	Naively Structured	True	Generative	200k	True
BERT	Direct Indexing	Naively Structured	True	Generative	10k	False
BERT	Summarization	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	True	Generative	10k	False
EncMoE-Decoder	Direct Indexing	Naively Structured	False	Generative	10k	False

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

Conclusion

RESULTS, PROBLEMS & FUTURE WORKS

Emanuele Rucci, Gianmarco Scarano

Sapienza Università di Roma

RESULTS

DECODING WITH

Greedy or Beam Search

Leads to 0% in all metrics: MAP, Recall@1000 & Hits@K

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

RESULTS

DISCRIMINATIVE APPROACH

LOSS

Cross-Entropy

TRAINING LOSS

0.001

Retrieval + Indexing

VALIDATION LOSS

>7.000

Retrieval

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

RESULTS

DIS

LOS

Cro

TRA

0.0

Retri

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

RESULTS

GENERATIVE APPROACH

LOSS

Cross-Entropy

VALIDATION LOSS

2.943

Retrieval

BEST MODEL

BERT

Lowest Loss for the Project

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

RESULTS

GE

BE

VAL

2.9

Retrie

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

RESULTS

Further relevant Metrics

15% - 20%

Percentage of valid DocIDs generated by the Models

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

PROBLEMS

Mismatch

Content of Doc VS Query content

If the answer to the Query is not in the indexed content, this should negatively influence the ability of retrieval.

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

PROBLEMS

Mismatch

Content of Doc VS Query content

If the answer to the Query is not in the indexed content, this should negatively influence the ability of retrieval.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In dictum, lectus et imperdiet viverra, justo justo congue nulla, ut aliquet velit justo sit amet lectus. Cras porta ex metus, at rutrum turpis varius quis. Sed eleifend tincidunt venenatis. Vestibulum ullamcorper laoreet lorem ut fringilla. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Phasellus velit ex, pharetra a dolor eget, ultrices imperdiet arcu. Mauris eu aliquet felis. Duis pulvinar pellentesque eros vulputate pretium. Aliquam faucibus turpis erat, et pretium dolor malesuada quis. Mauris dictum venenatis nisl, non ultricies nunc blandit sit amet. Aliquam maximus dignissim ipsum non euismod. Praesent efficitur est sit amet lacus volutpat, feugiat ullamcorper dolor accumsan.

Nullam interdum turpis

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

PROBLEMS

GENERALIZATION

Poor ability to generalize

Learning **Indexing** seems “effective”, but **Retrieval** depends on generalization capabilities.

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

PROBLEMS

GENERALIZATION

Poor ability to generalize

Learning **Indexing** seems “effective”, but **Retrieval** depends on generalization capabilities.

1. Setacciare la farina e versarla in una ciotola insieme a zucchero, un pizzico di sale, lievito e scorza grattugiata di limone.
2. Unire ora gli ingredienti umidi, ovvero un uovo, un tuorlo ed il burro fuso nel microonde oppure a bagnomaria. Mescolare per bene, prima con il cucchiaio poi a mano, aggiungendo eventualmente un paio di cucchiai di acqua per rendere più facile la lavorazione.
3. Accendere il forno a 180°C.
4. Foderate 2 teglie da forno con la carta oleata.
5. Spolverizzare la spianatoia con un po' di farina e stendere la pasta con l'aiuto di ...

How many grams of sugar are needed?

What’s the amount of sugar I have to buy for this recipe?

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA

PROBLEMS

Quantization

Limits the precision

By lowering the precision during computations, the performances are affected negatively.

Emanuele Rucci, Gianmarco Scarano

Sapienza
Università di Roma

FUTURE WORKS

Improvements & Enhancements

Mixtral-8×7B

What is the effect of a MoE-Decoder?

MISTRAL AI_

LoRA-The-Explorer

Optimized Quantization strategy w.r.t standard LoRA[1]

Microsoft's AICI

Constraints the Decoder to output DocIDs[2]

[1]: https://arxiv.org/pdf/2402.16828.pdf
[2]: https://github.com/microsoft/aici

Emanuele Rucci, Gianmarco Scarano

SAPIENZA UNIVERSITÀ DI ROMA

THANK YOU!

QUESTIONS?

Emanuele Rucci, Gianmarco Scarano

SAPIENZA
UNIVERSITÀ DI ROMA