Natural Language Processing

Encoder-based language models: BERT

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING UNIVERSITY

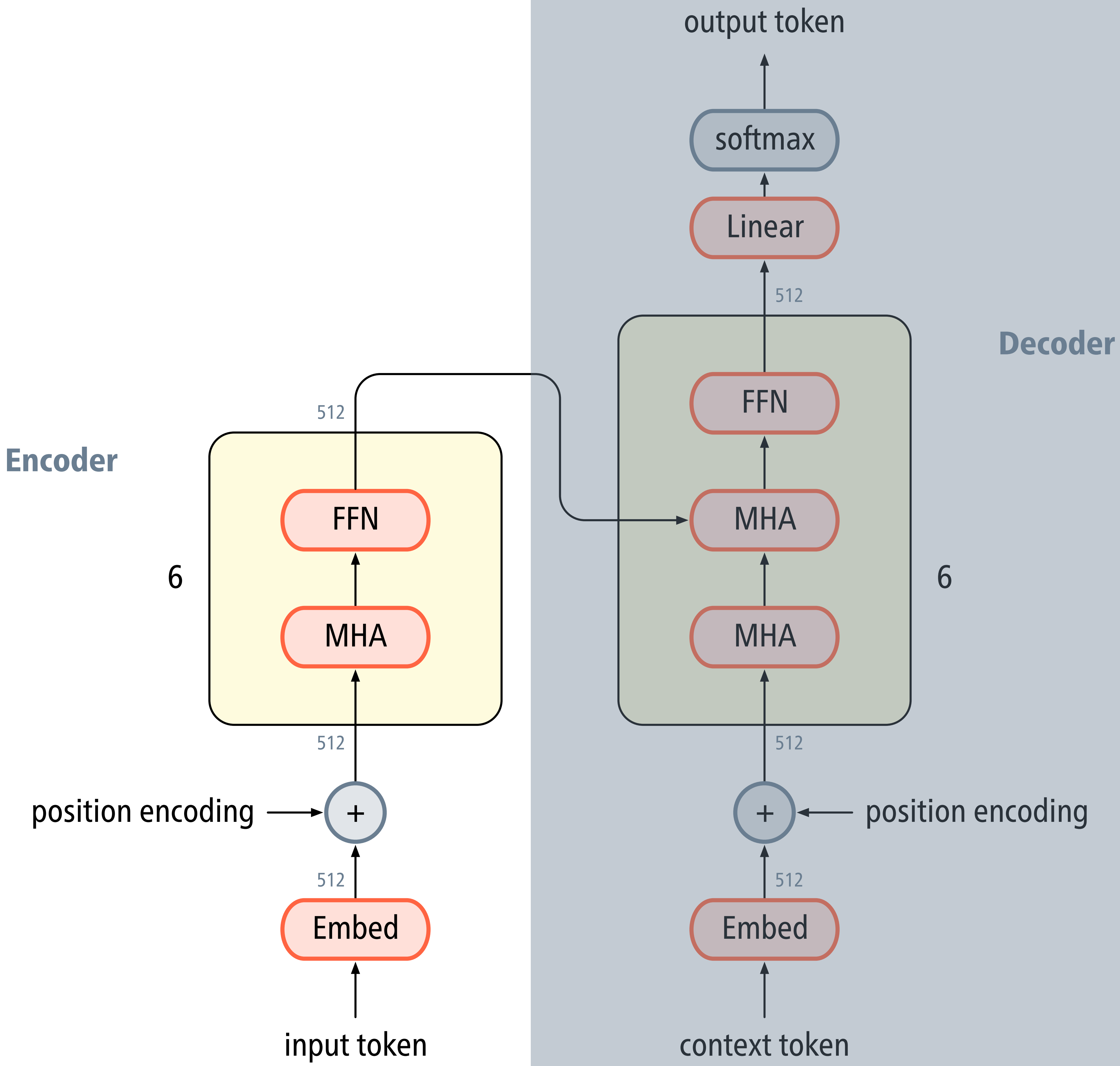
BERT

- The acronym **BERT** stands for “Bidirectional Encoder Representations from Transformers”.

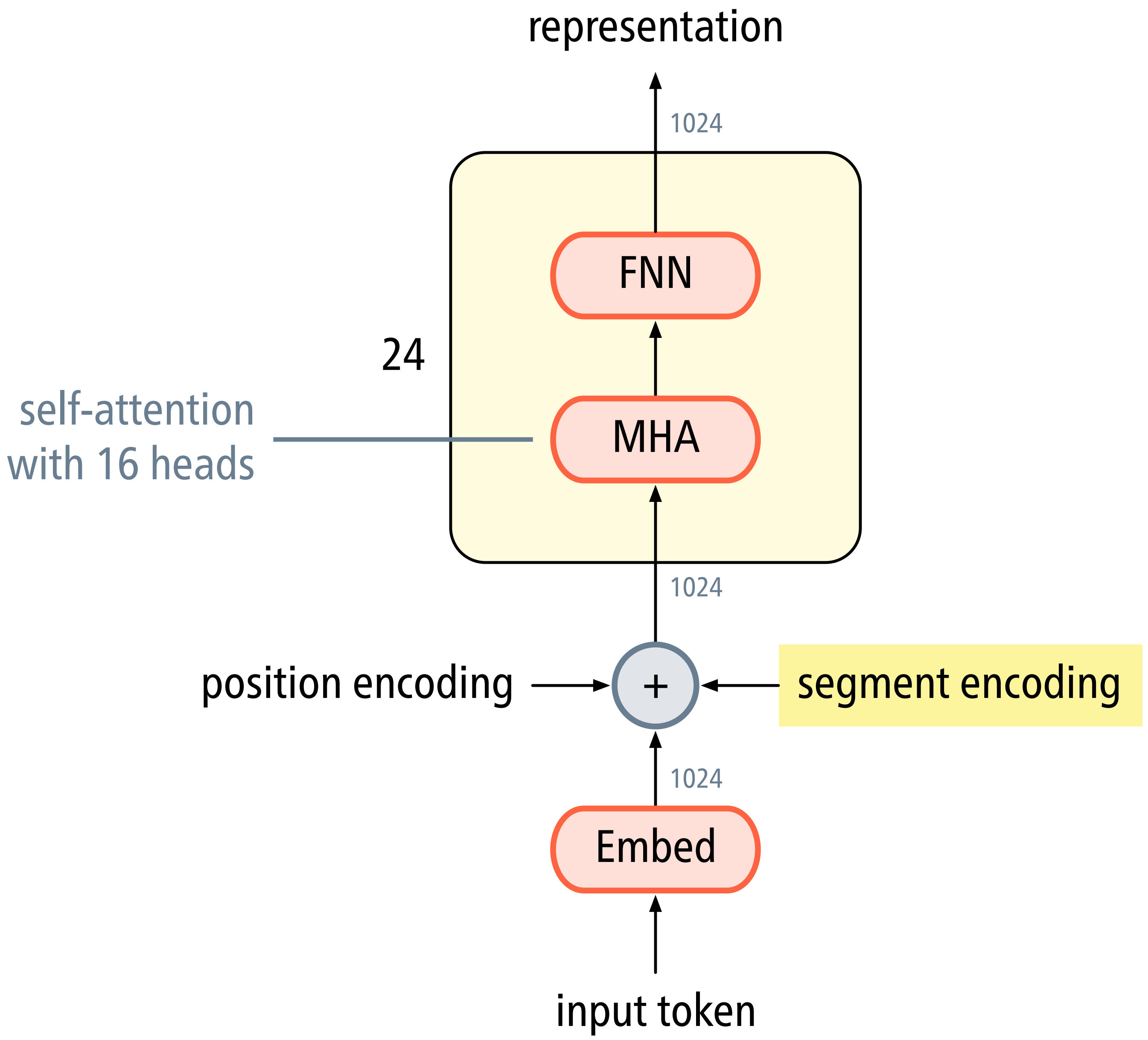
- As an encoder, BERT can learn token representations that are conditioned on the complete input sequence.

  non-directional

Devlin et al. (2019) | Muppet character image from The Muppet Wiki


output token
softmax
Linear
512
Decoder
FFN
MHA
MHA
6
512
position encoding
+
512
Embed
context token
Encoder
512
FFN
MHA
6
512
position encoding
+
512
Embed
input token

BERT (large model)



Devlin et al. (2019)

Model statistics

	base	large
number of dimensions	768	1024
number of encoder blocks	12	24
number of attention heads	12	16
number of parameters	**110 M**	**340 M**

Devlin et al. (2019)

Pre-training tasks

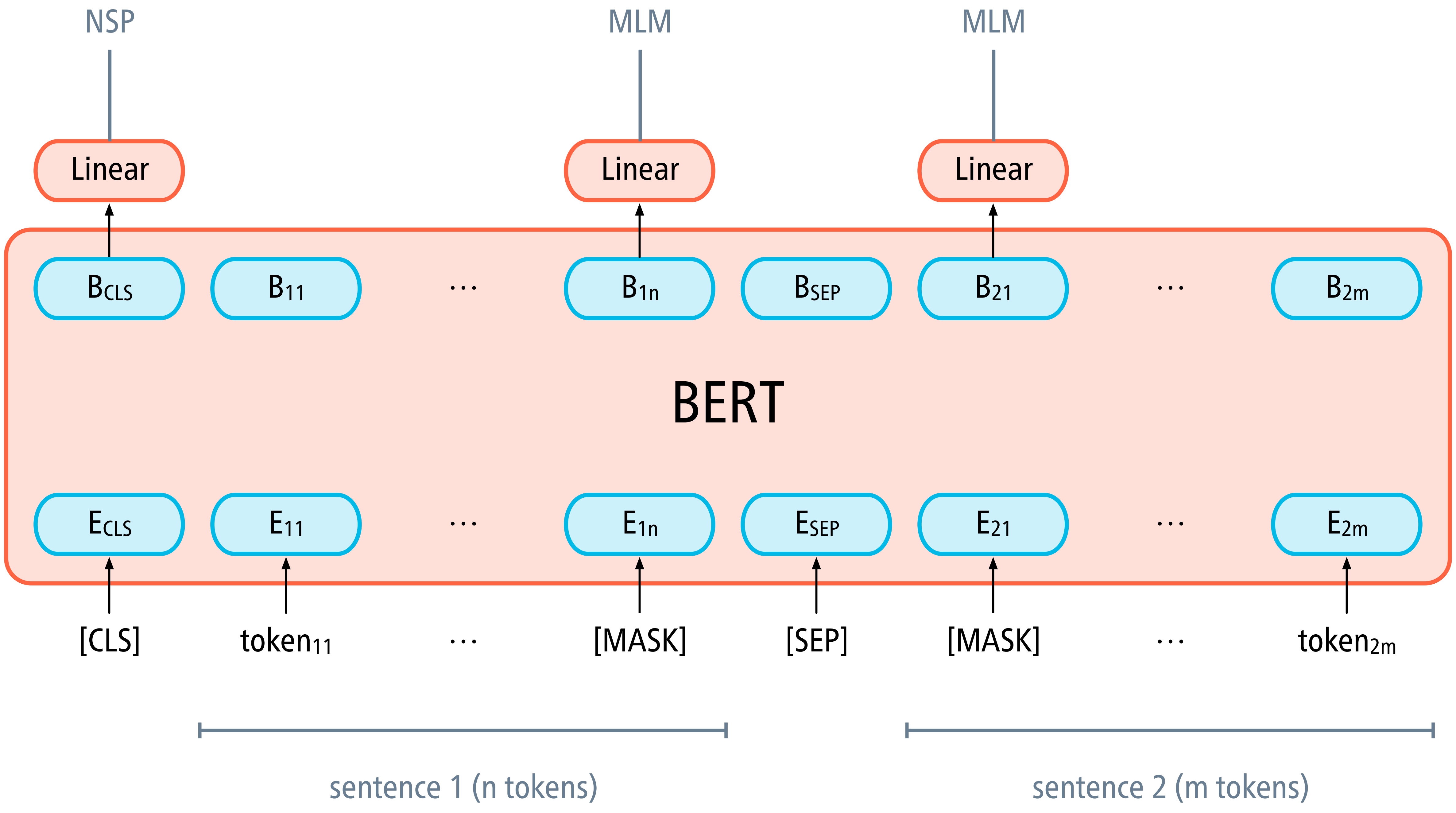
- **Masked Language Modelling (MLM)**

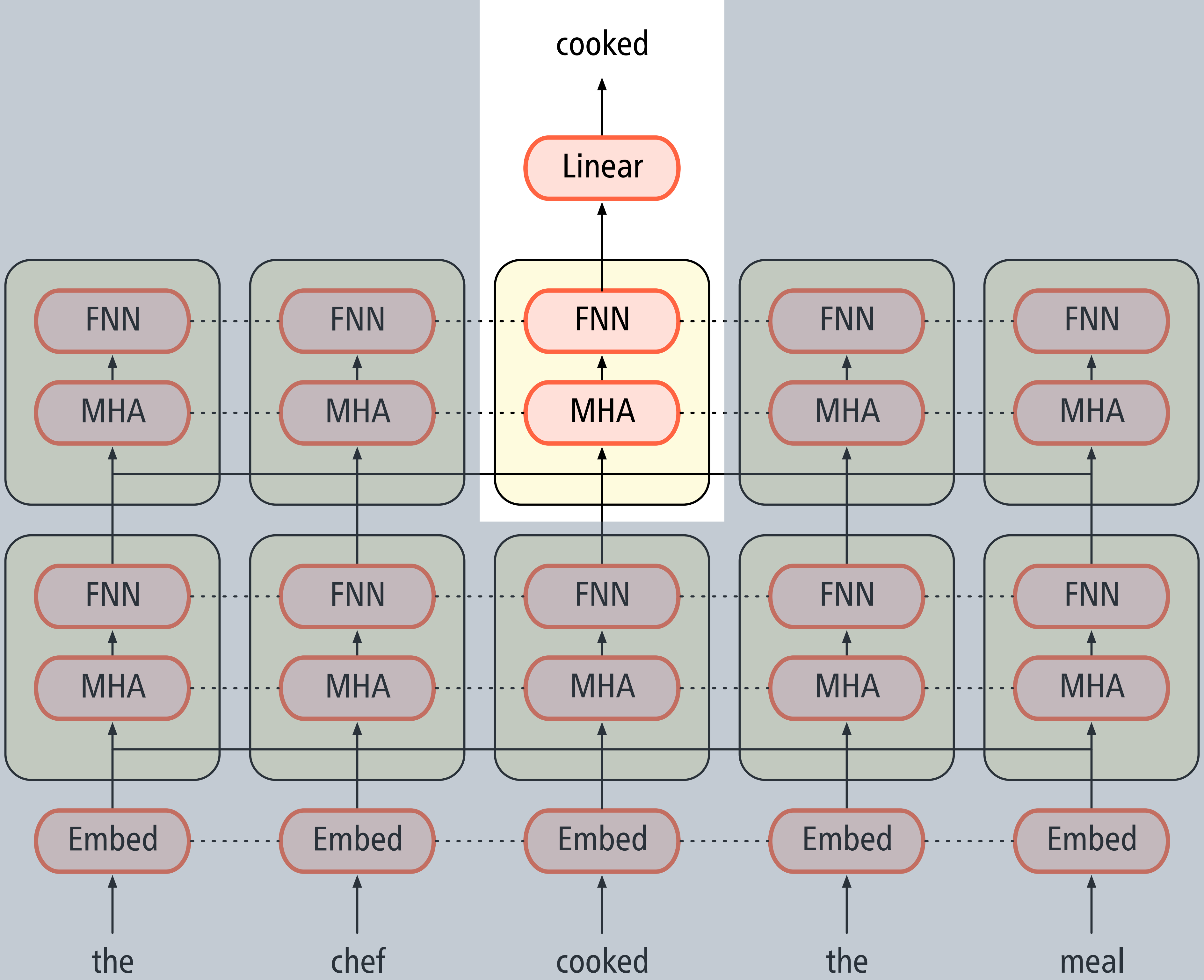
  Tokens are masked out at random. The model is trained to predict the masked-out tokens.

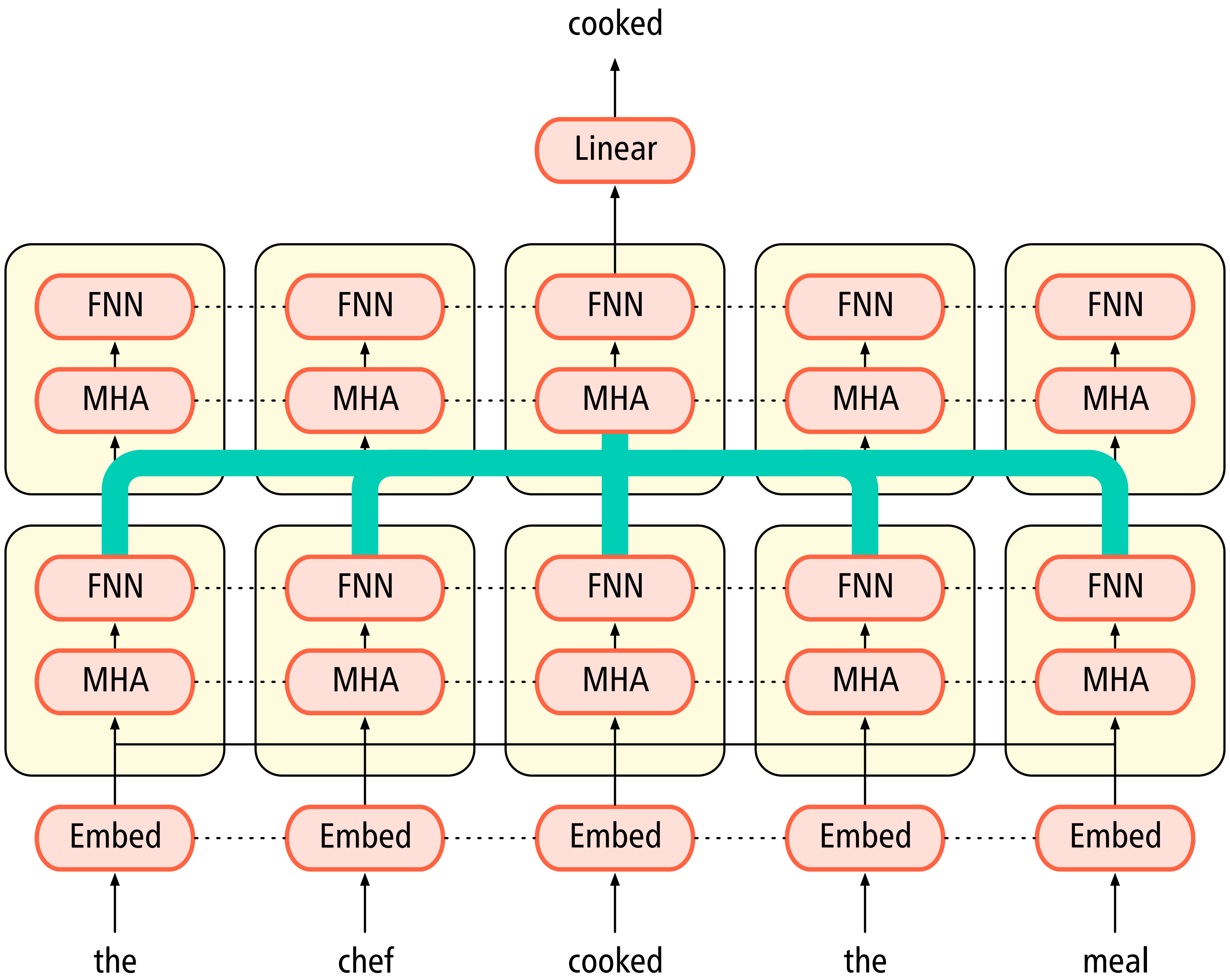
- **Next Sentence Prediction (NSP)**

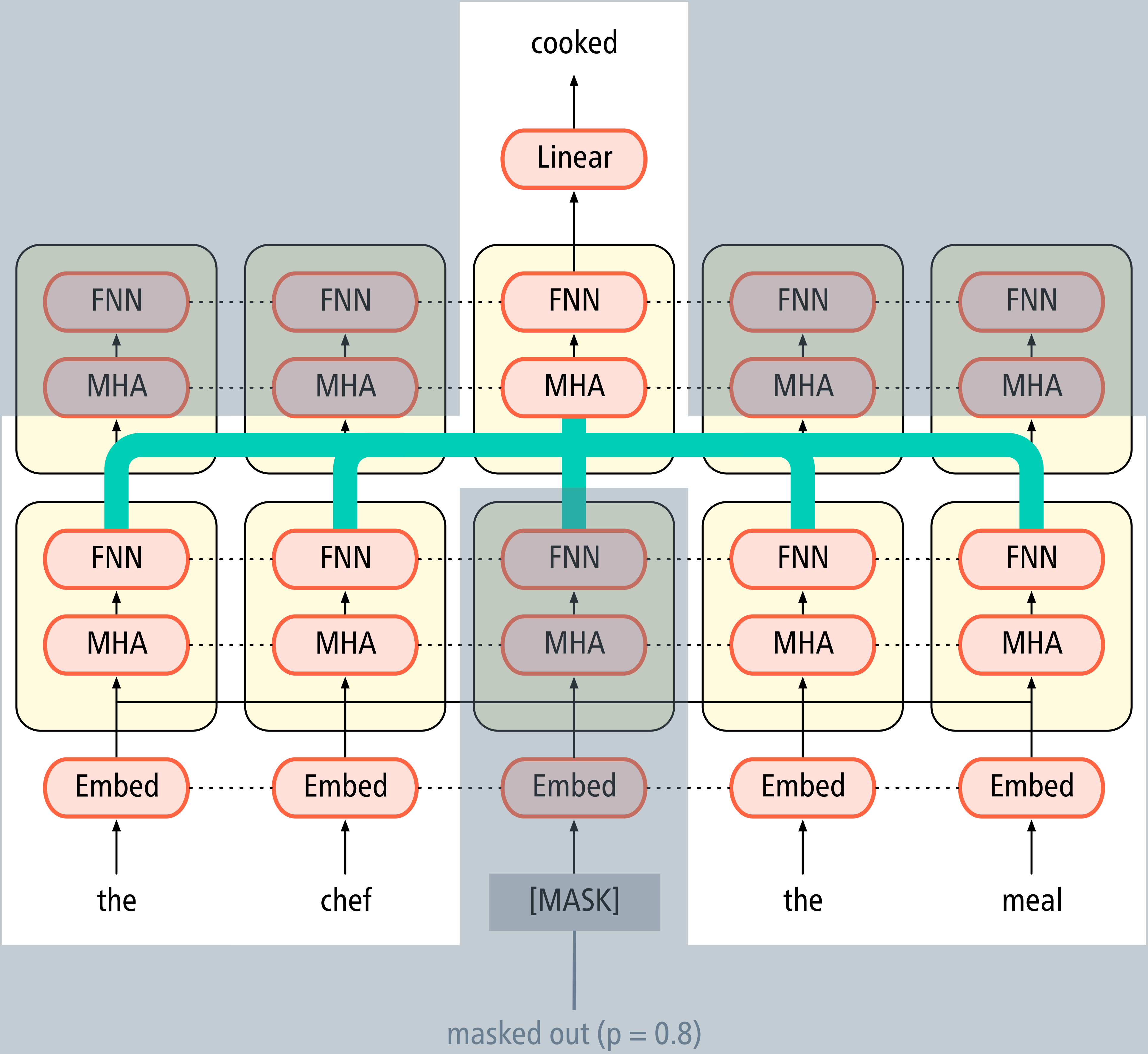
  The model is trained to predict whether two randomly sampled sentences are adjacent in the training data.

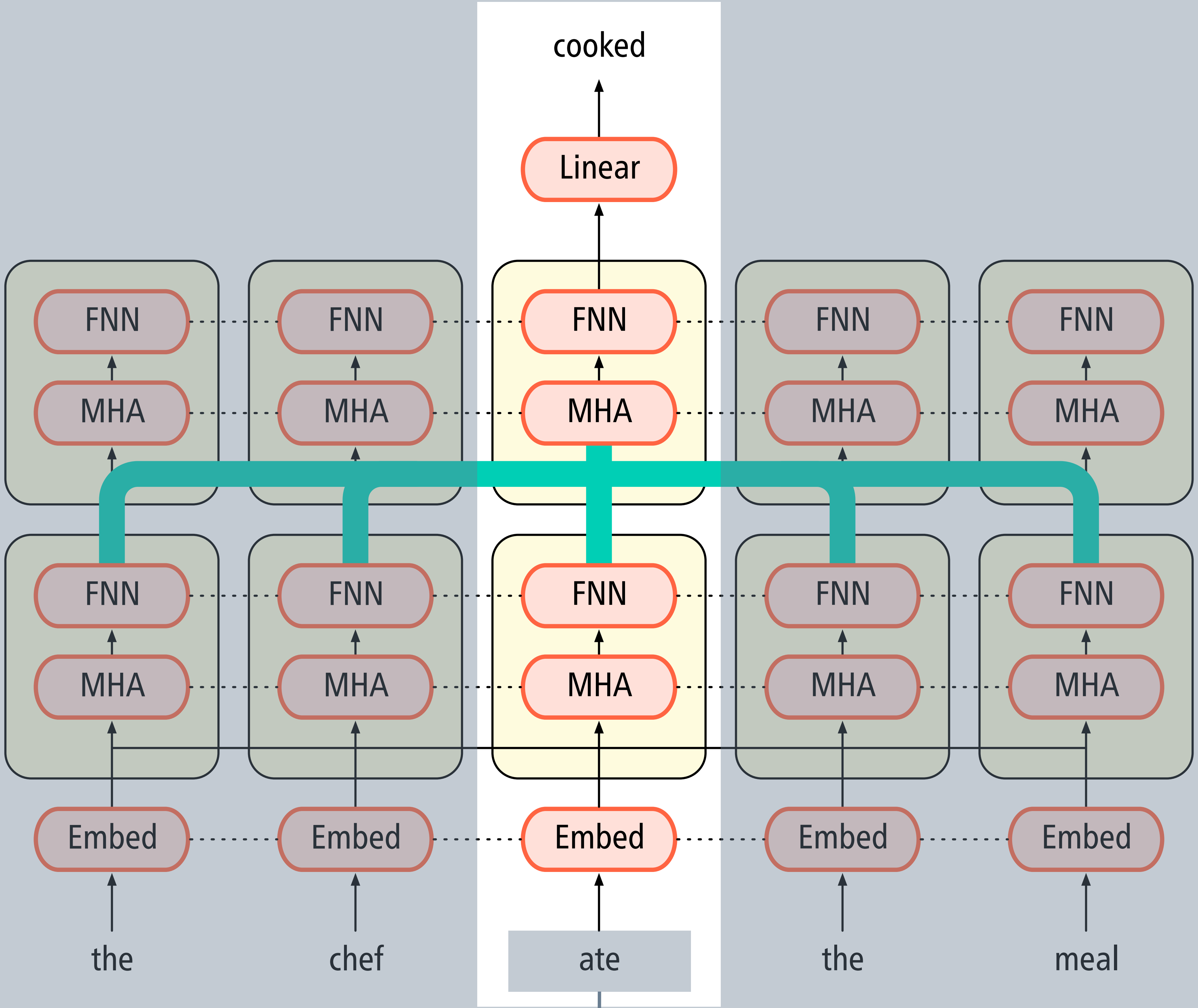
Pre-training with MLM and NSP



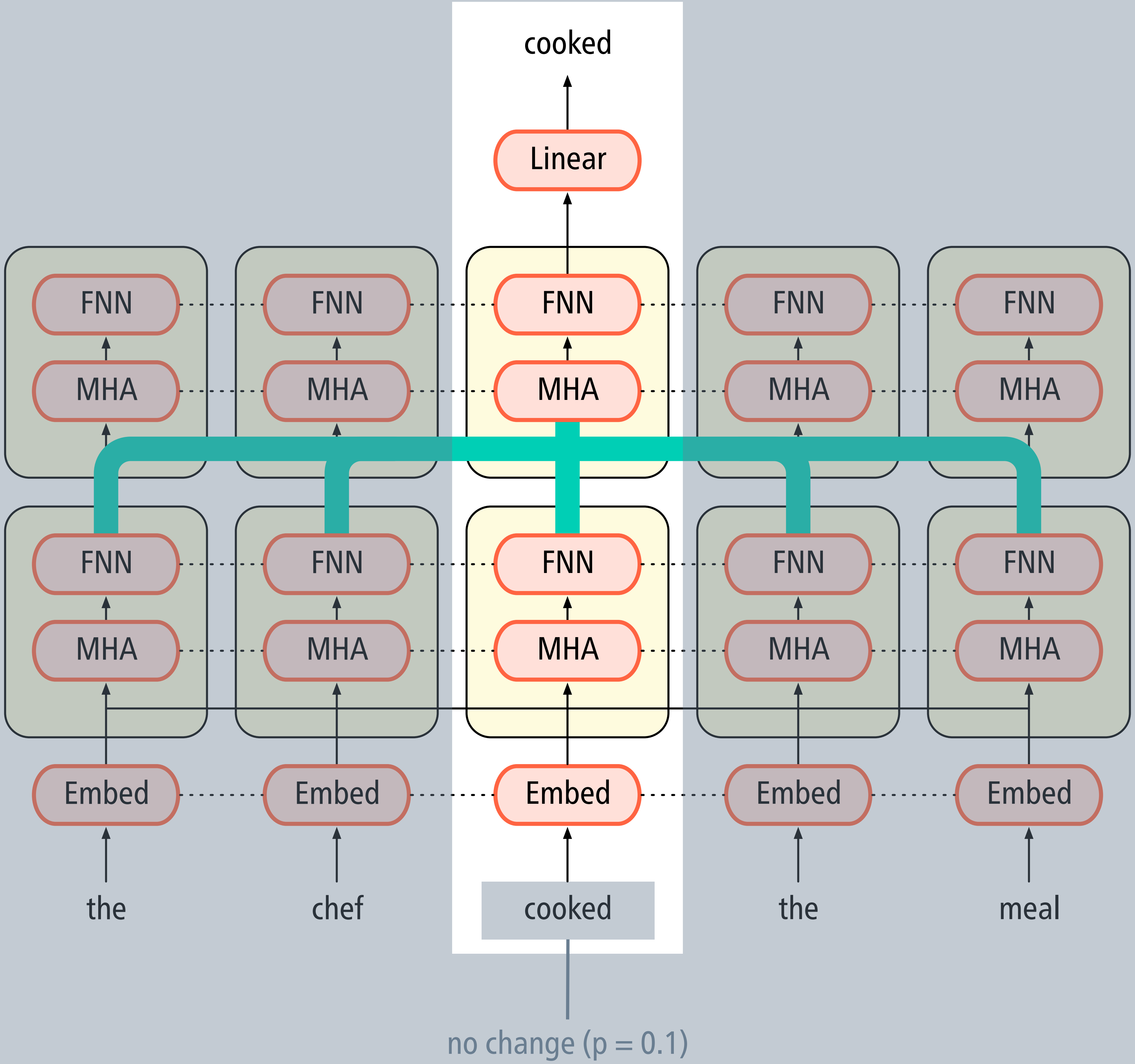

cooked
Linear
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
Embed
Embed
Embed
Embed
Embed
the
chef
cooked
the
meal


cooked
Linear
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
Embed
Embed
Embed
Embed
Embed
the
chef
cooked
the
meal

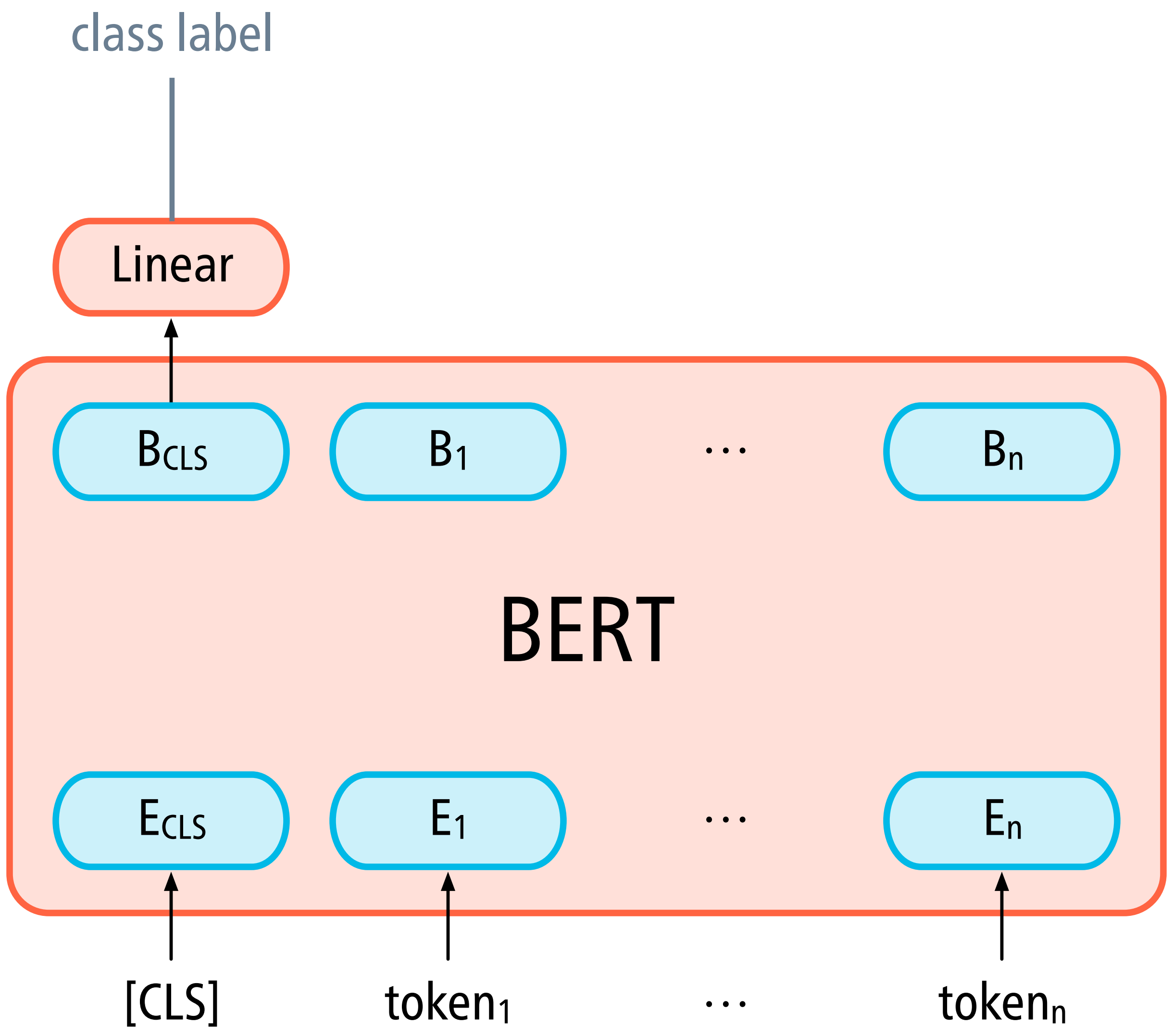

cooked
Linear
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
Embed
Embed
Embed
Embed
Embed
the
chef
[MASK]
the
meal
masked out (p = 0.8)



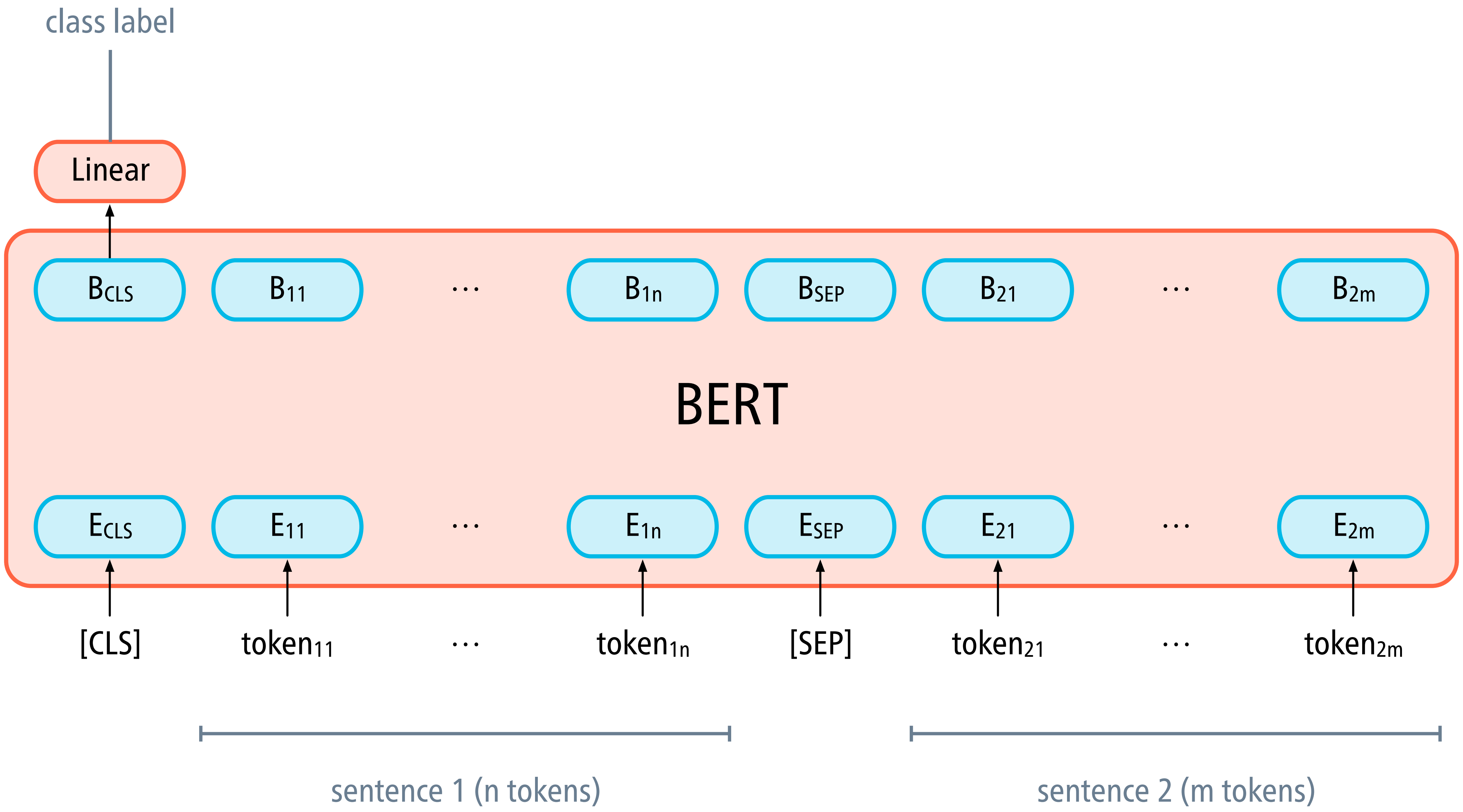
replaced with random word (p = 0.1)


cooked
Linear
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
FNN
FNN
FNN
FNN
FNN
MHA
MHA
MHA
MHA
MHA
Embed
Embed
Embed
Embed
Embed
the
chef
cooked
the
meal
no change (p = 0.1)

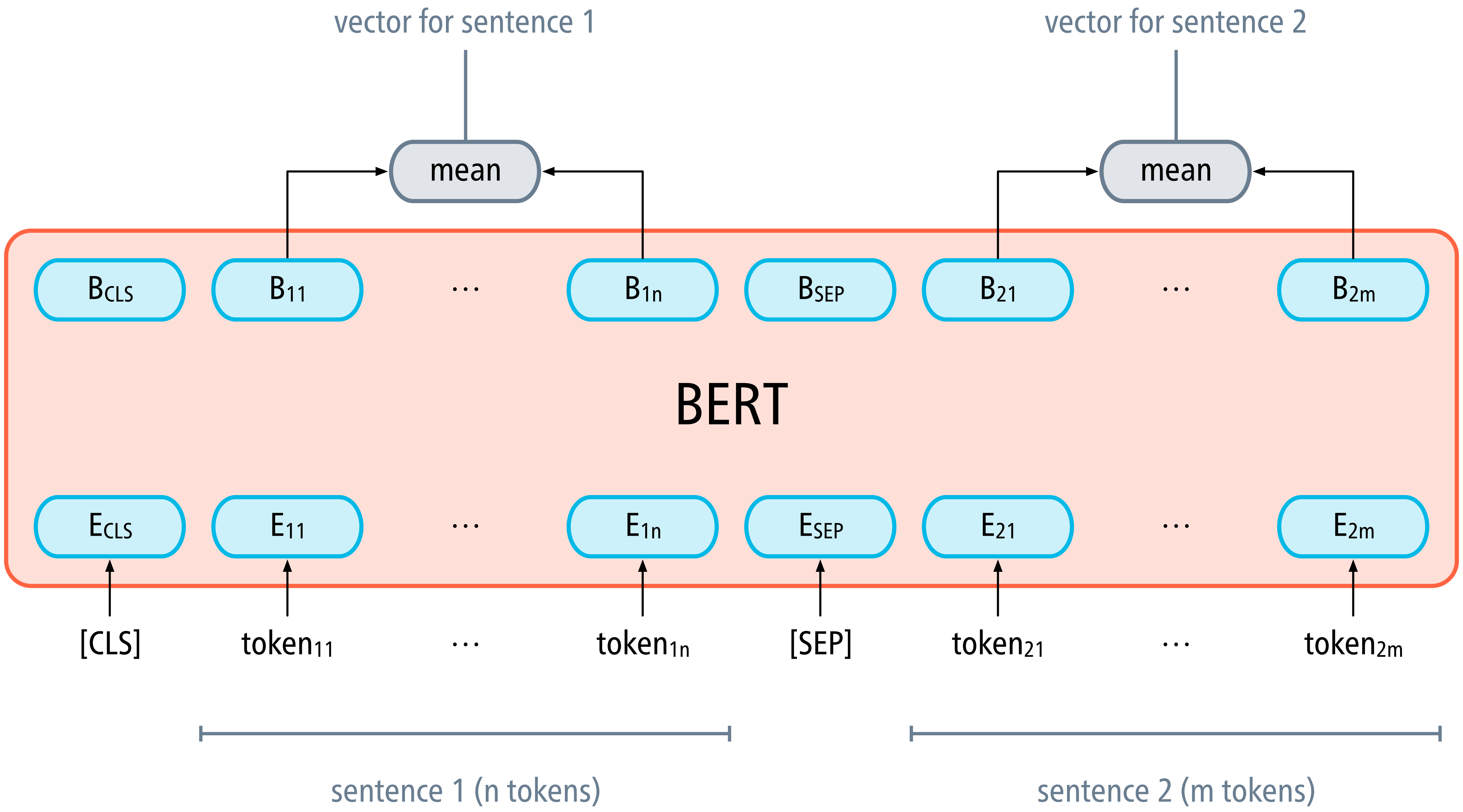
Fine-tuning on a single-sentence classification task



Fine-tuning on a sentence-pair classification task



Fine-tuning on a sentence-pair similarity task



Performance on the GLUE benchmark

	GLUE
ELMo + Attention	71.0
Previous state-of-the-art	74.0
BERT (base)	79.6
BERT (large)	**82.1**

GLUE test results, scored by the evaluation server | Devlin et al. (2019)

BERT-like models

- RoBERTa uses an improved recipe for pre-training and a significantly larger data set.

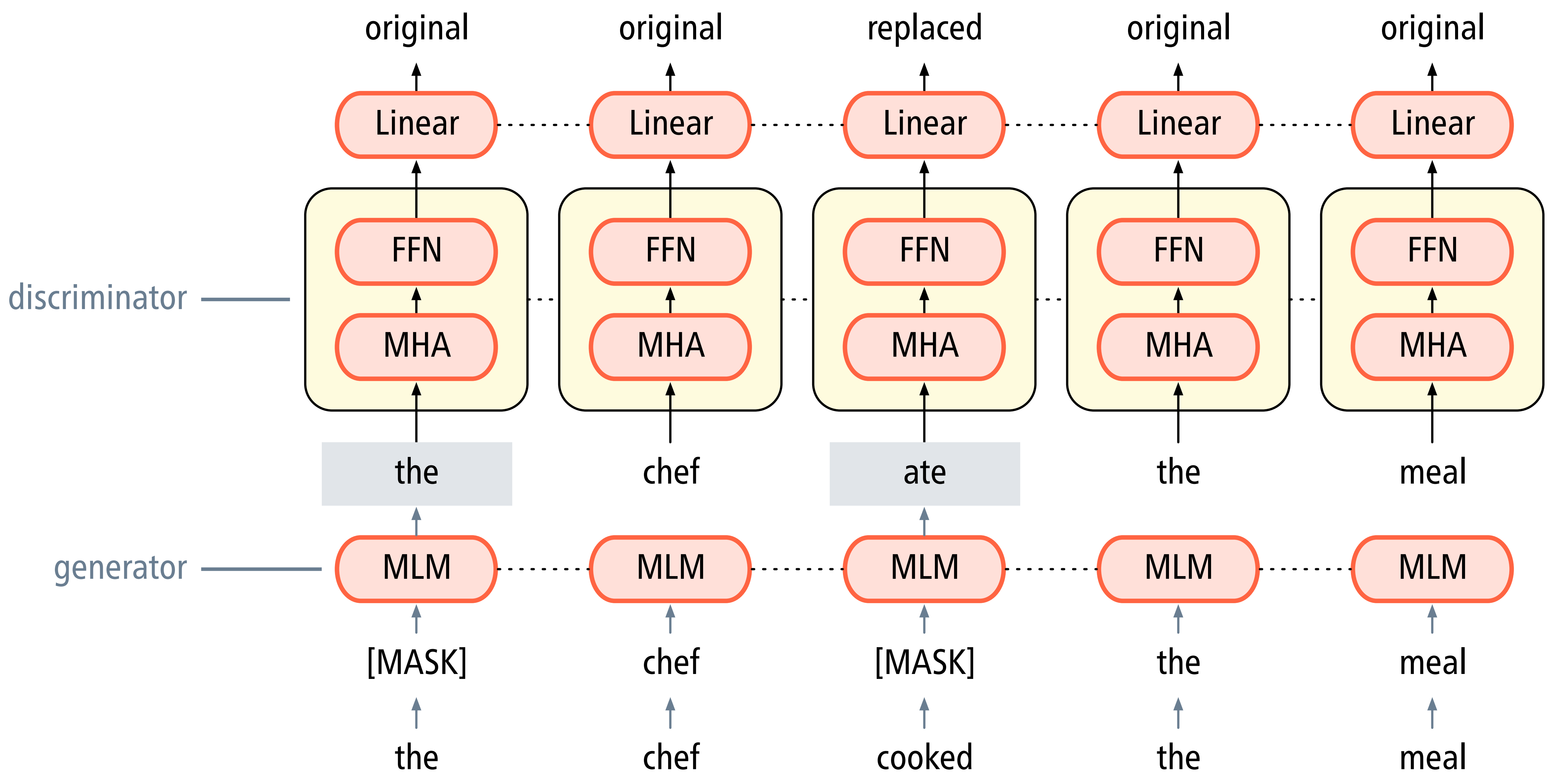
  Liu et al. (2019)

- ALBERT and DistilBERT are models with reduced training time and model size, respectively.

  Lan et al. (2019), Sanh et al. (2019)

- Many pre-trained BERT-like and other transformer models are available via Hugging Face.

ELECTRA: Pre-training via replaced token detection



Clark et al. (2020)

Effectiveness of replaced token detection

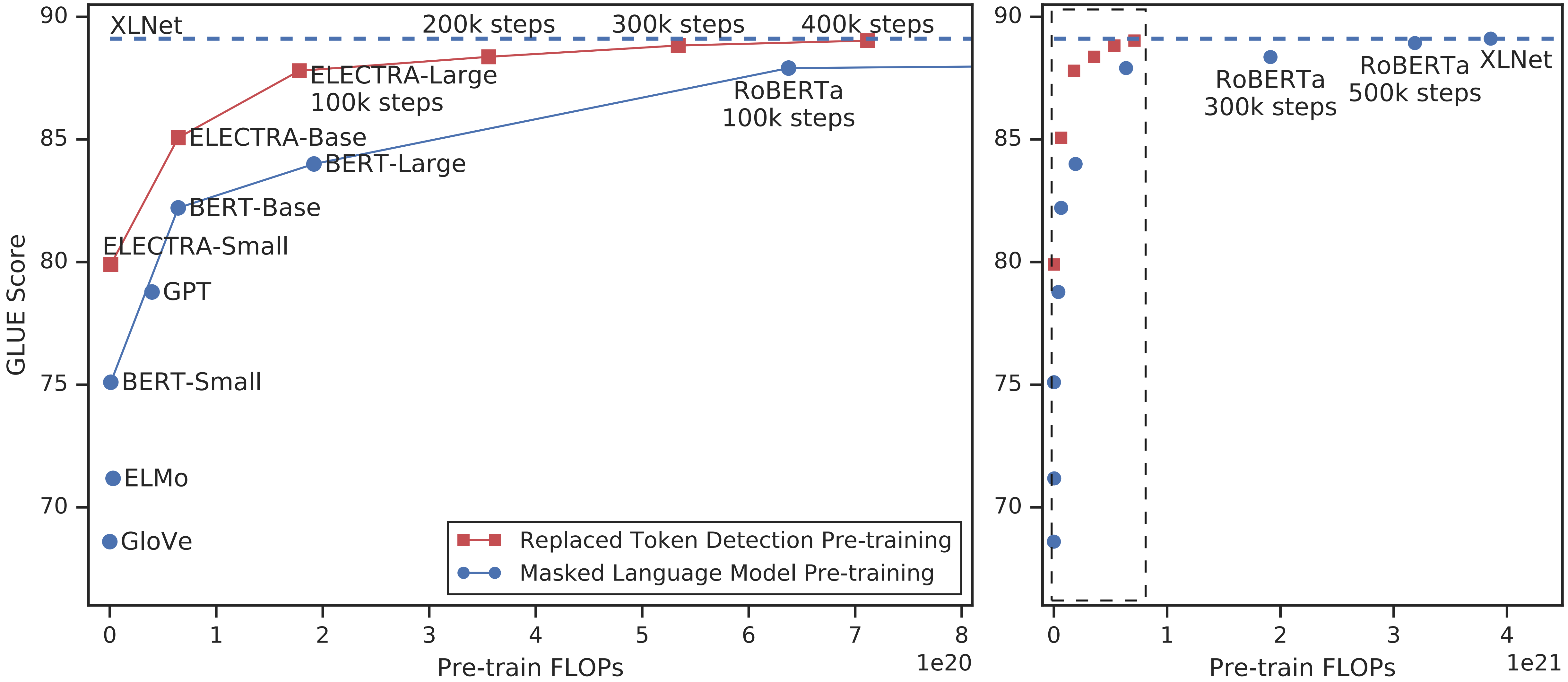


Figure 1: Replaced token detection pre-training consistently outperforms masked language model pre-training given the same compute budget. The left figure is a zoomed-in view of the dashed box.

Clark et al. (2020)