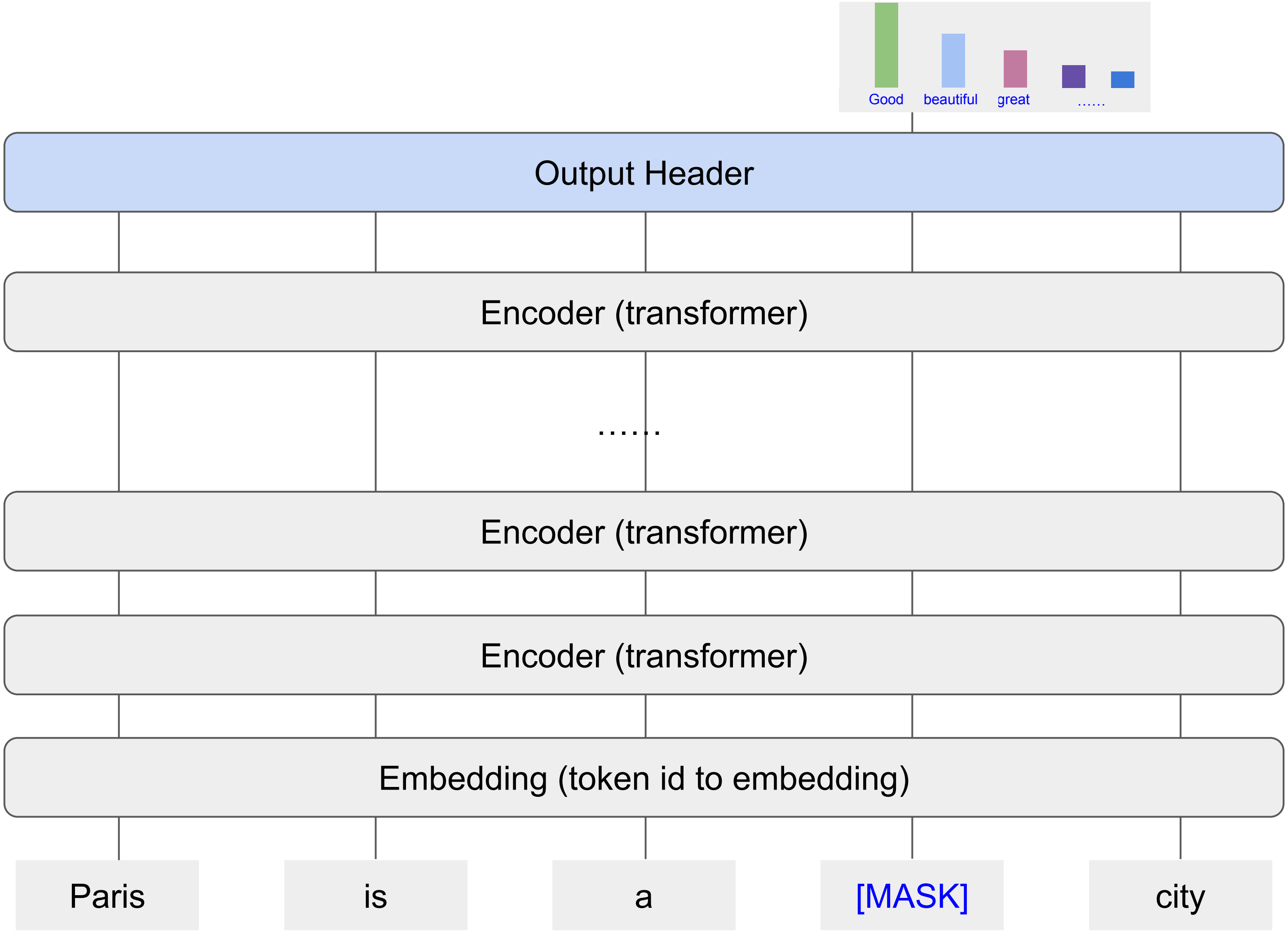
BERT Masked Token Prediction



Good

beautiful

great

……

Output Header

Encoder (transformer)

……

Encoder (transformer)

Encoder (transformer)

Embedding (token id to embedding)

Paris

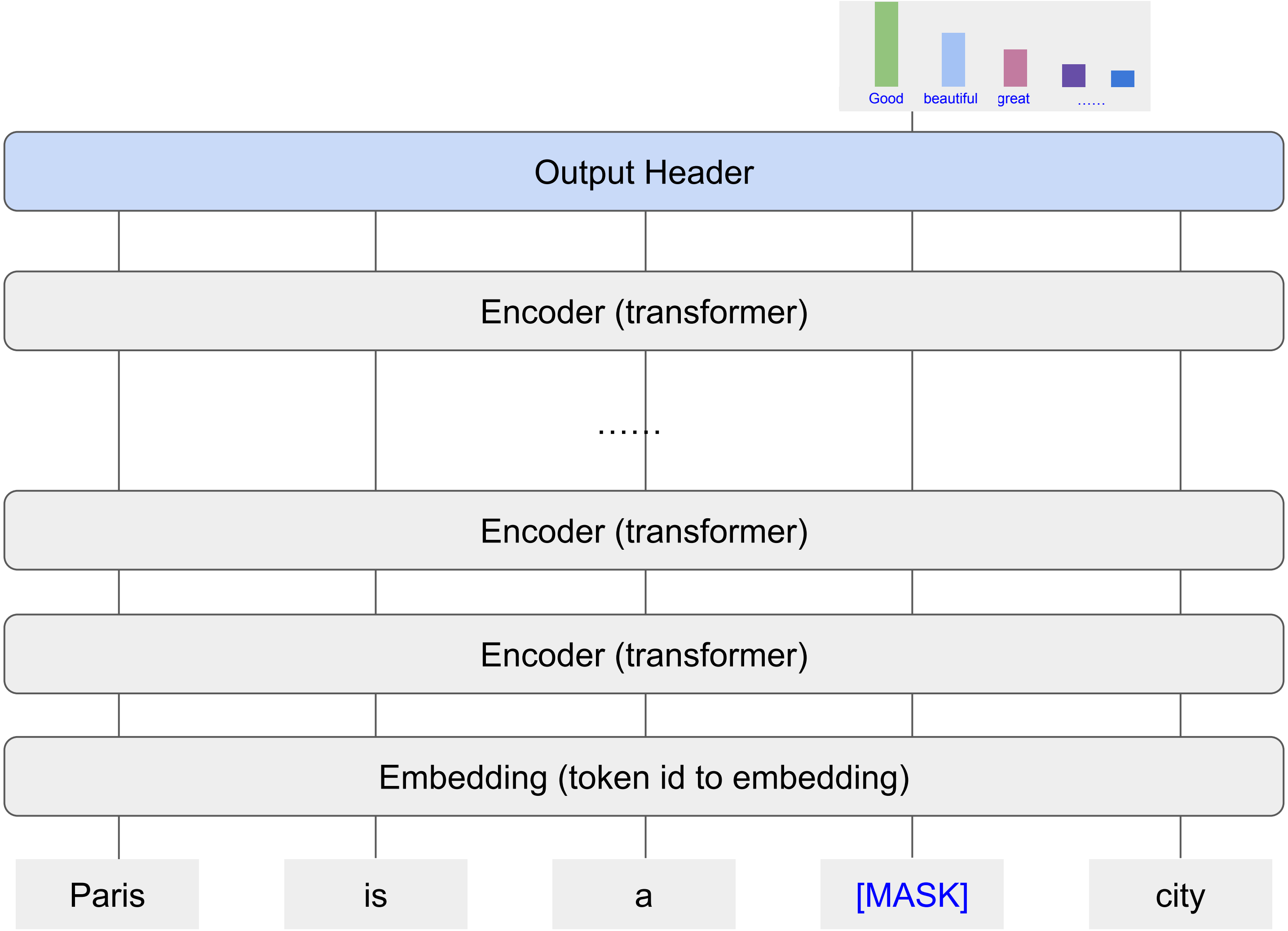
is

a

[MASK]

city

BERT Masked Token Prediction



Output Header

Encoder (transformer)

......

Encoder (transformer)

Encoder (transformer)

Embedding (token id to embedding)

Paris

is

a

[MASK]

city

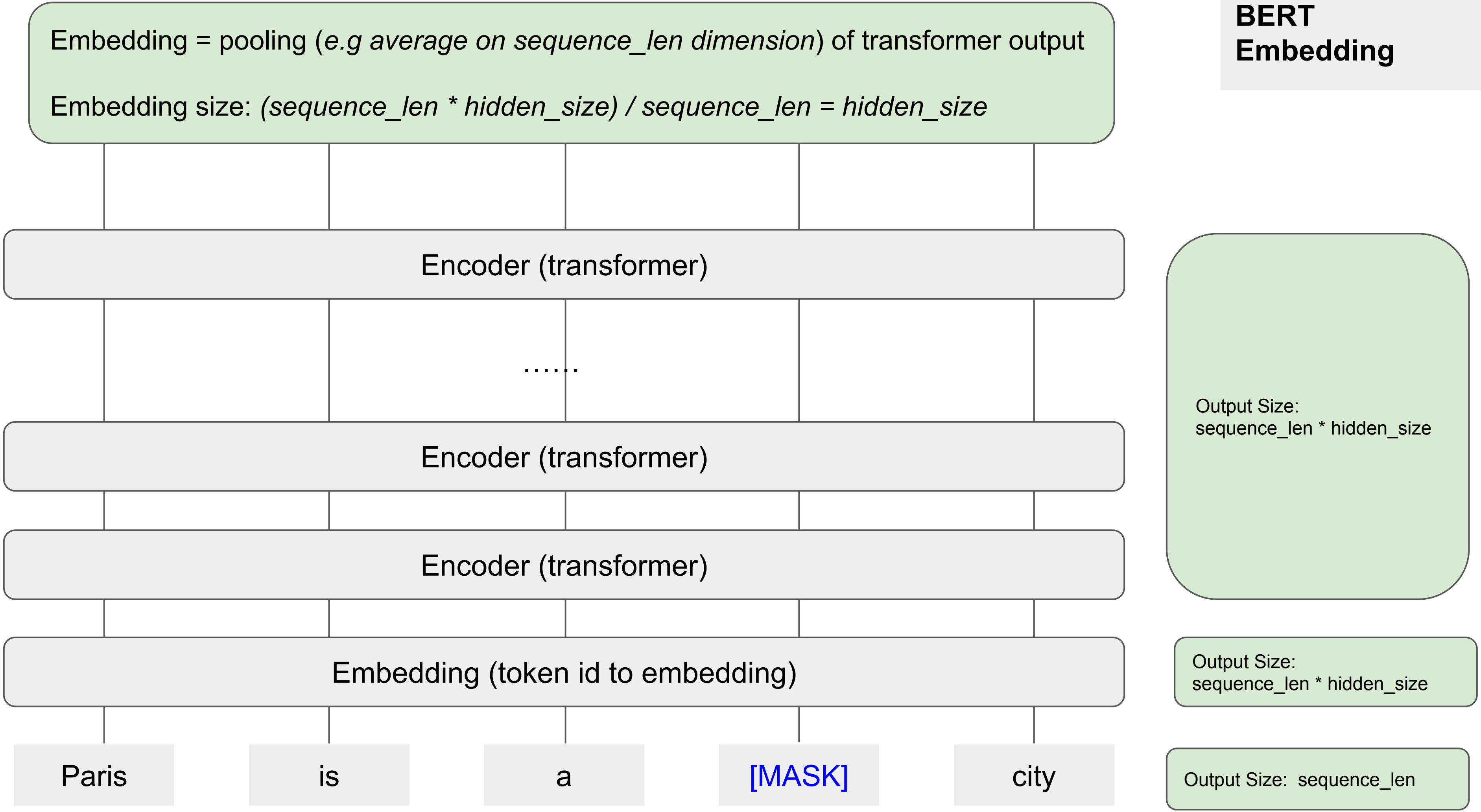
Output Size: vocab_size
(each word in vocab has a
probability, sum is 100%)

Output Size:
sequence_len * hidden_size

Output Size:
sequence_len * hidden_size

Output Size: sequence_len

BERT Embedding



Embedding = pooling (*e.g average on sequence_len dimension*) of transformer output

Embedding size: *(sequence_len * hidden_size) / sequence_len = hidden_size*

Encoder (transformer)

……

Encoder (transformer)

Encoder (transformer)

Embedding (token id to embedding)

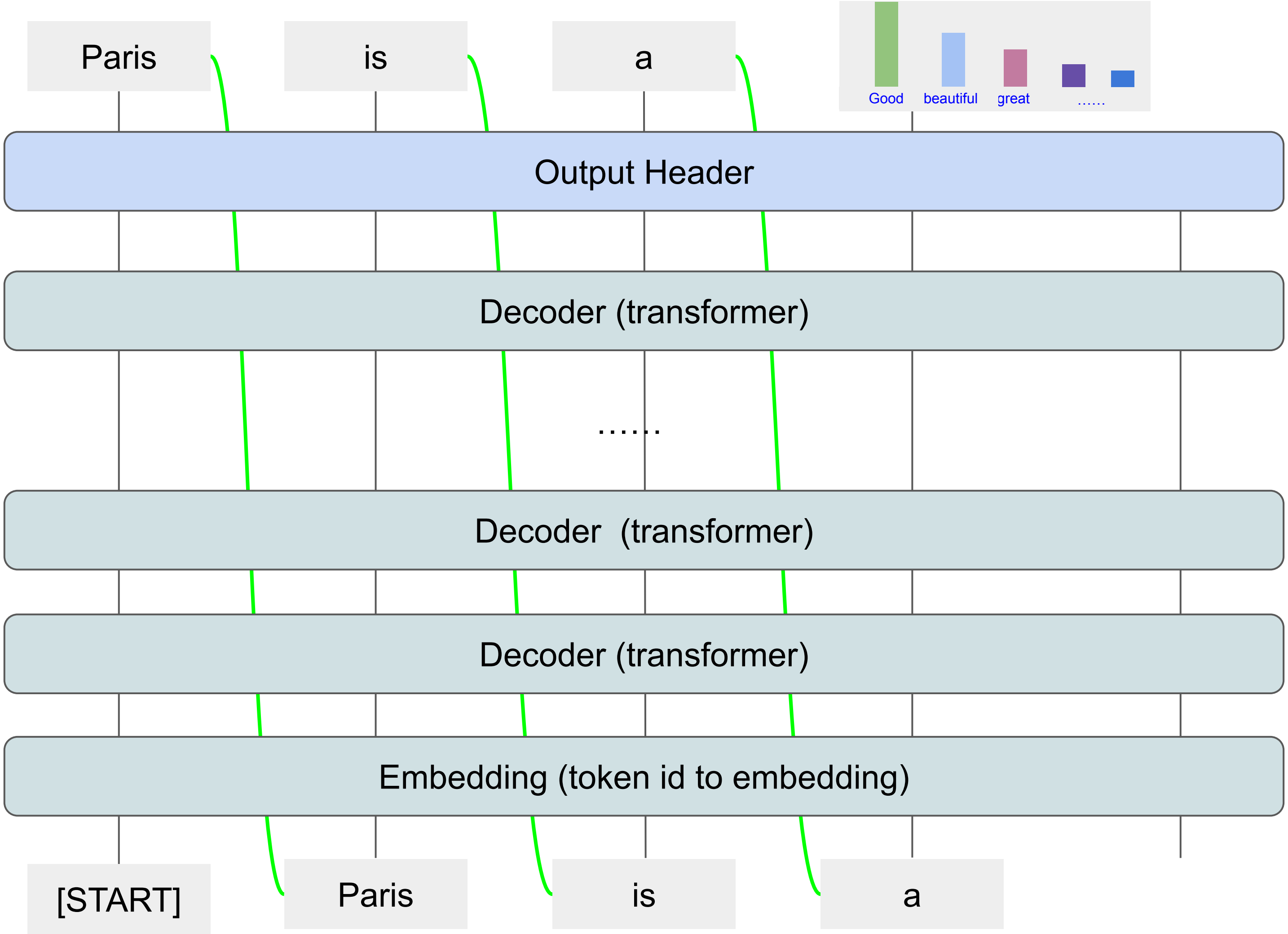
Paris is a [MASK] city

Output Size:
sequence_len * hidden_size

Output Size:
sequence_len * hidden_size

Output Size: sequence_len

GPT Next Token Prediction



Paris

is

a

Output Header

Decoder (transformer)

......

Decoder (transformer)

Decoder (transformer)

Embedding (token id to embedding)

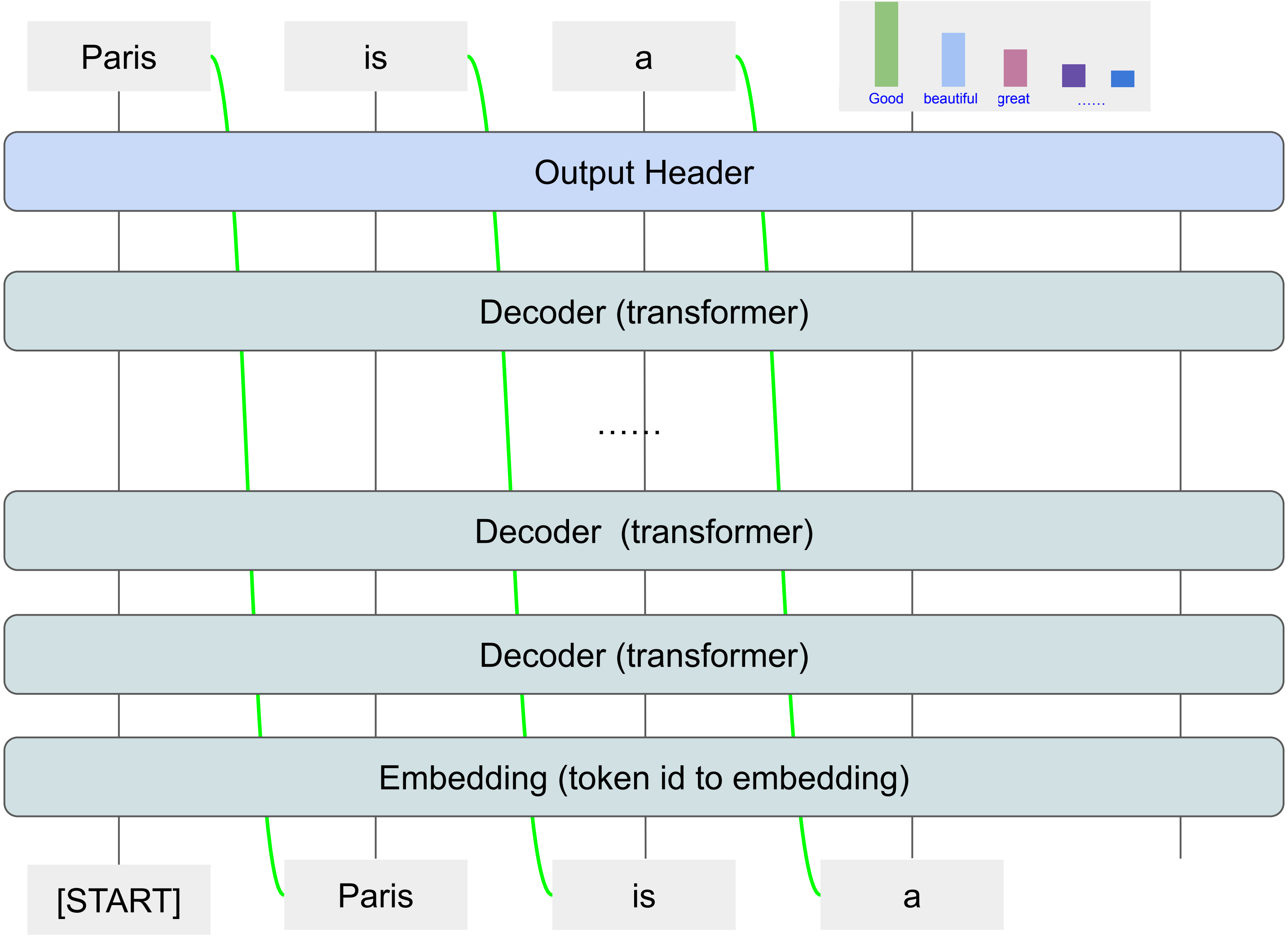
[START]

Paris

is

a

GPT Next Token Prediction



Paris

is

a

Output Header

Decoder (transformer)

……

Decoder (transformer)

Decoder (transformer)

Embedding (token id to embedding)

[START]

Paris

is

a

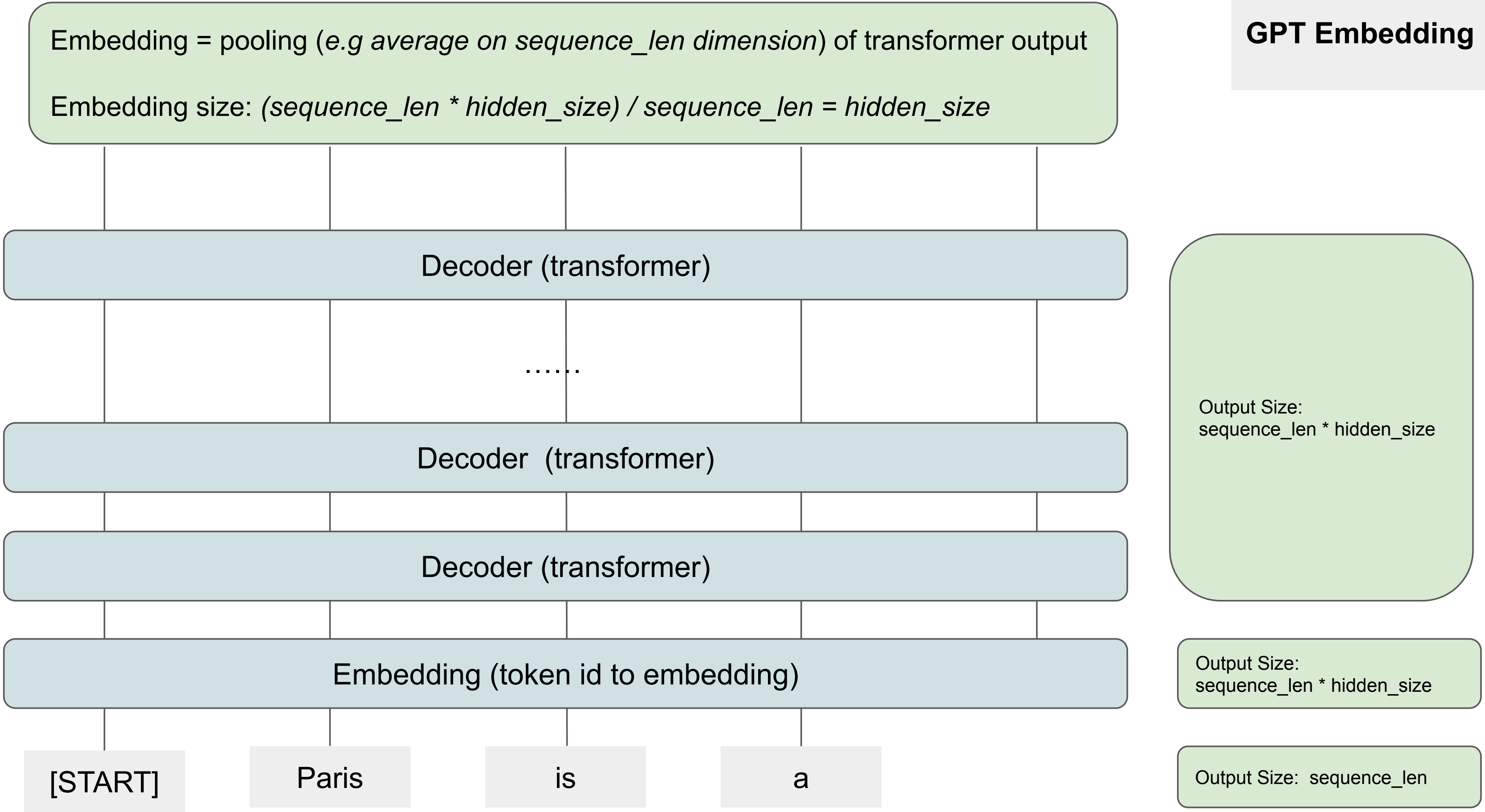
Output Size: vocab_size
(each word in vocab has a
probability, sum is 100%)

Output Size:
sequence_len * hidden_size

Output Size:
sequence_len * hidden_size

Output Size: sequence_len

GPT Embedding



Embedding = pooling (*e.g average on sequence_len dimension*) of transformer output

Embedding size: *(sequence_len * hidden_size) / sequence_len = hidden_size*

Decoder (transformer)

……

Decoder (transformer)

Decoder (transformer)

Embedding (token id to embedding)

[START]

Paris

is

a

Output Size:
sequence_len * hidden_size

Output Size:
sequence_len * hidden_size

Output Size: sequence_len