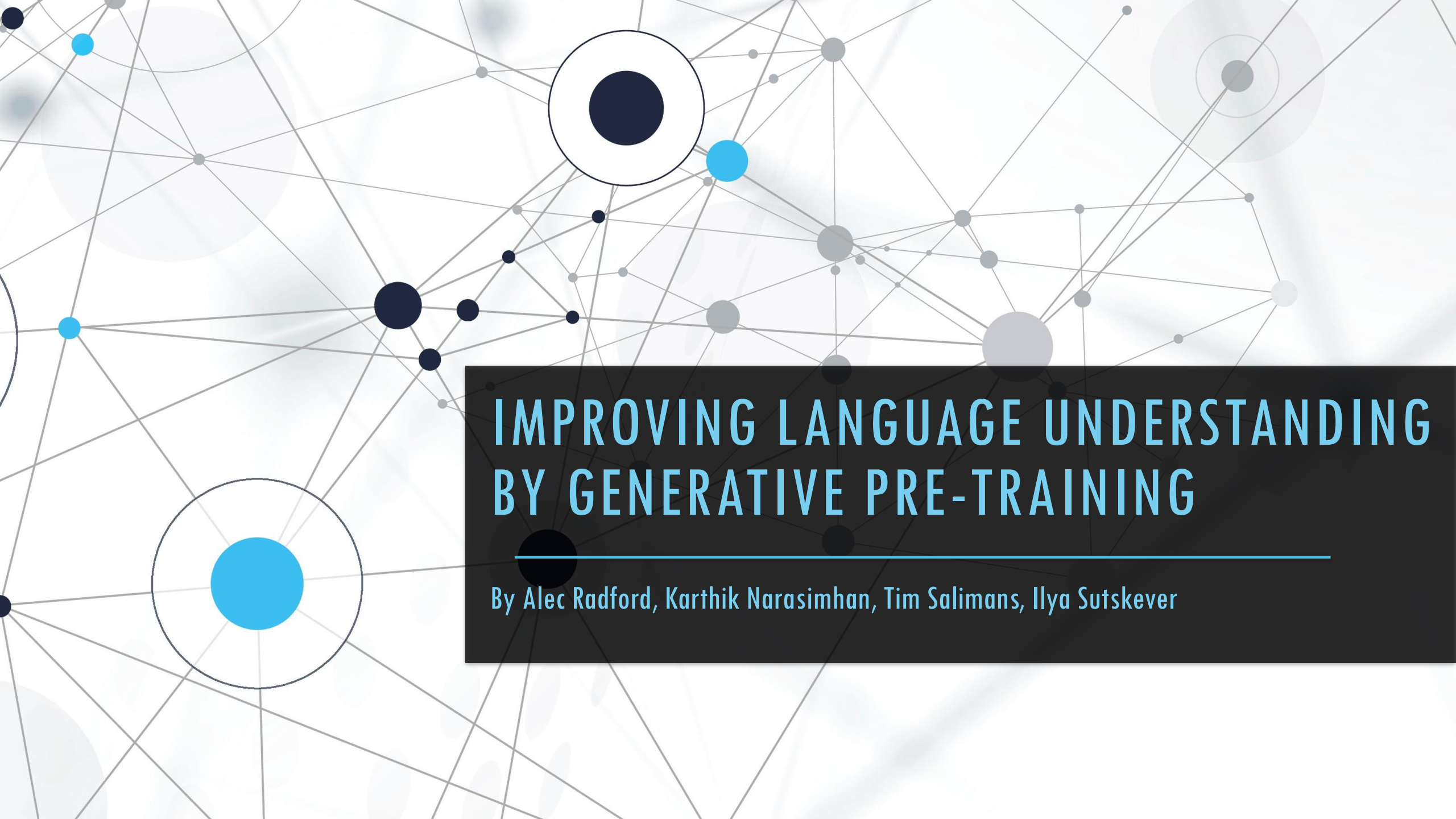


IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING

By Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

GPT-1

By Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

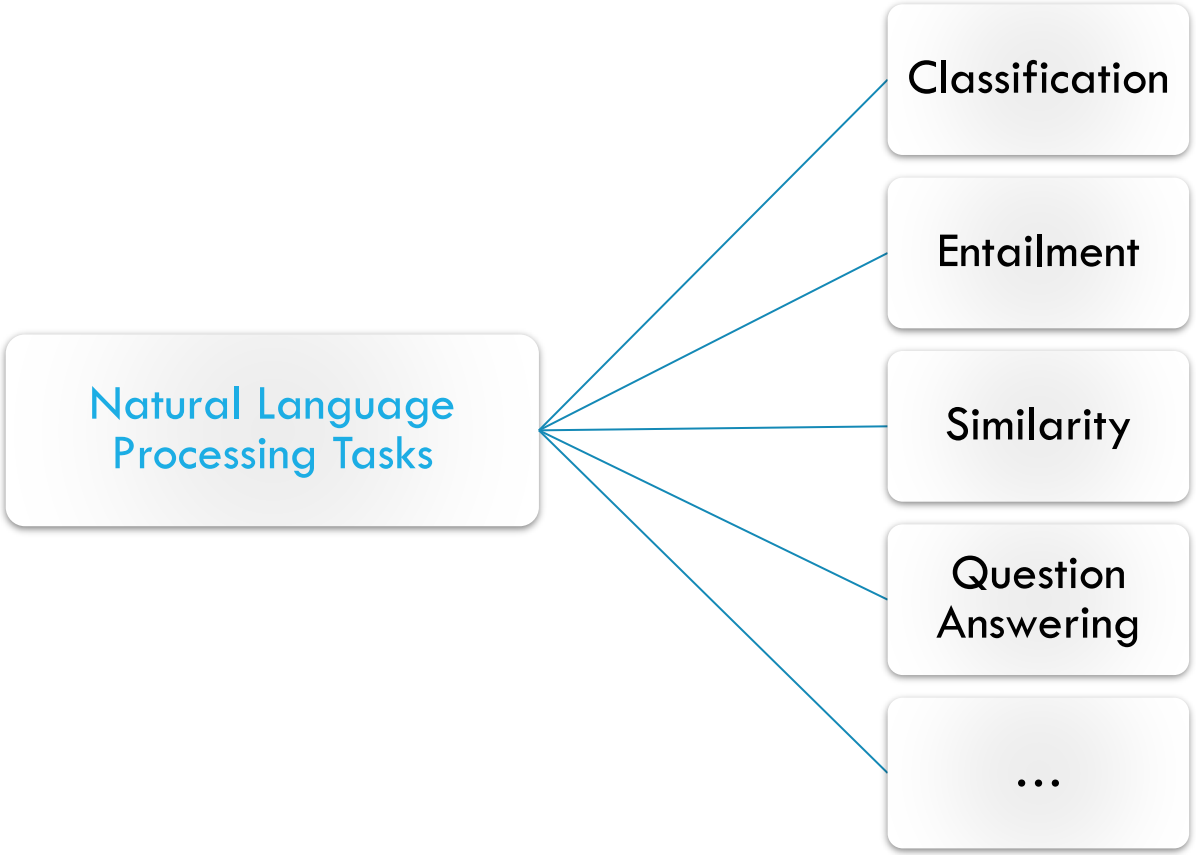
OVERVIEW

1) Motivation
2) Model Architecture
3) Framework
4) Evaluation
5) Conclusion
6) Discussion

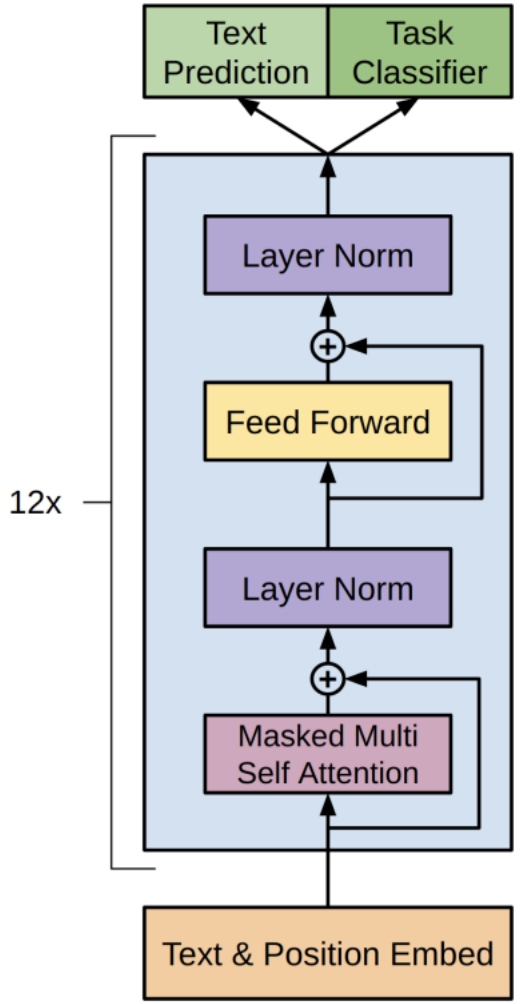
1) MOTIVATION

Answer the question.	„Which country does Berlin belong to?“
Is one sentence part of the other?	1) „She plays with Mary tomorrow.“ 2) „She plays with a girl tomorrow.“
Positive or negative?	„This performance was stunning!“

1) MOTIVATION

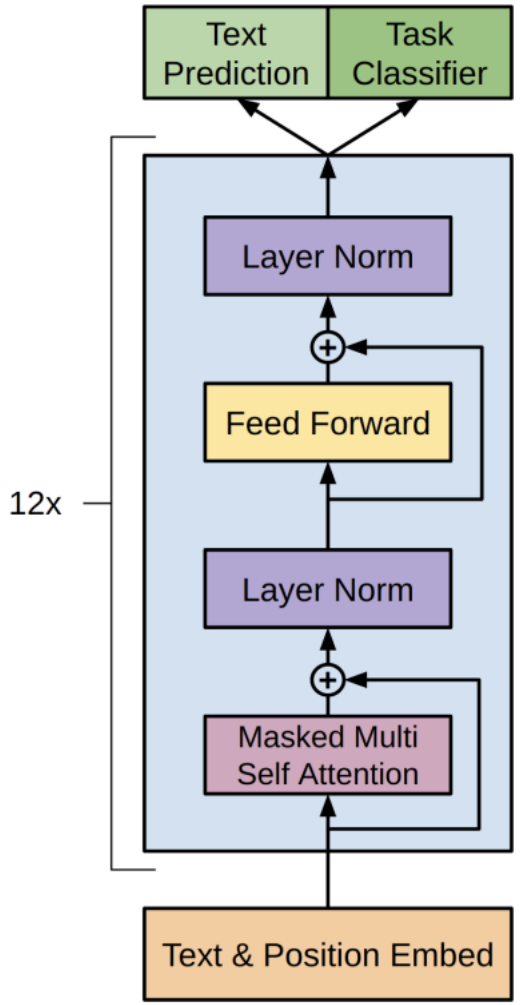


2) MODEL ARCHITECTURE



*3

2) MODEL ARCHITECTURE



*3

INPUT EMBEDDING

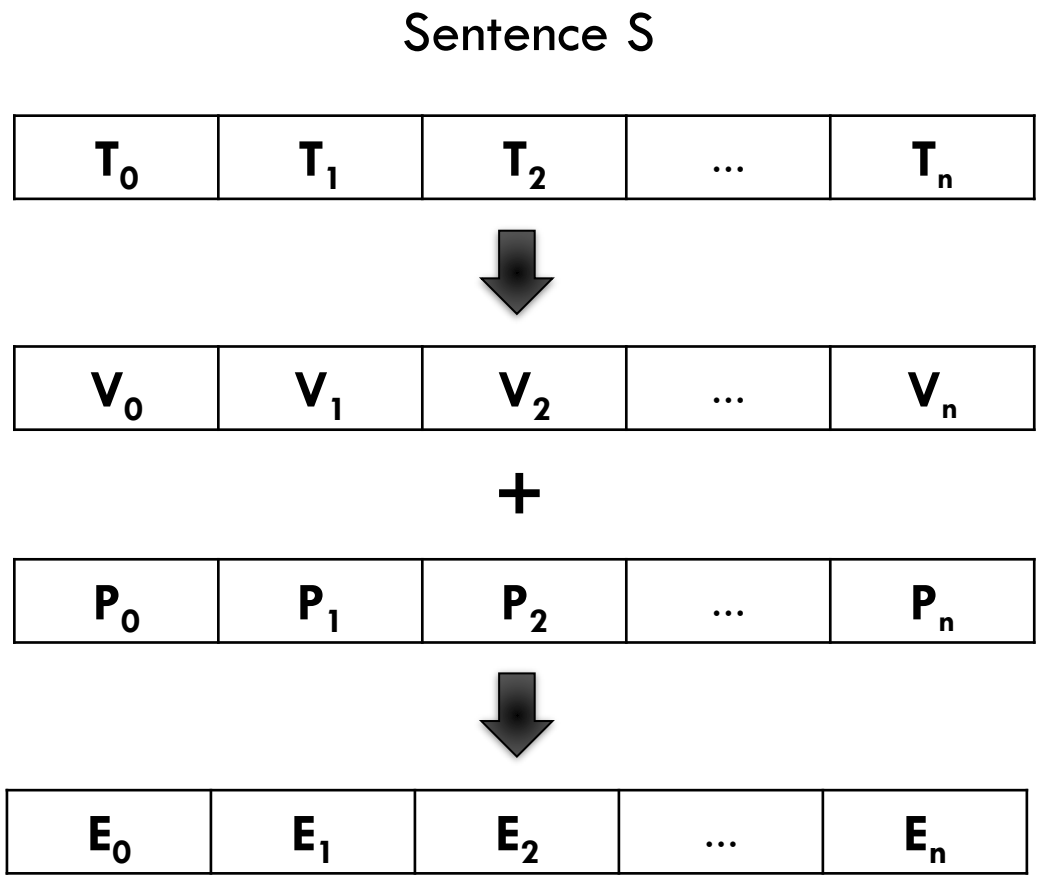
Input

Tokenization via Byte-Pair-Encodings

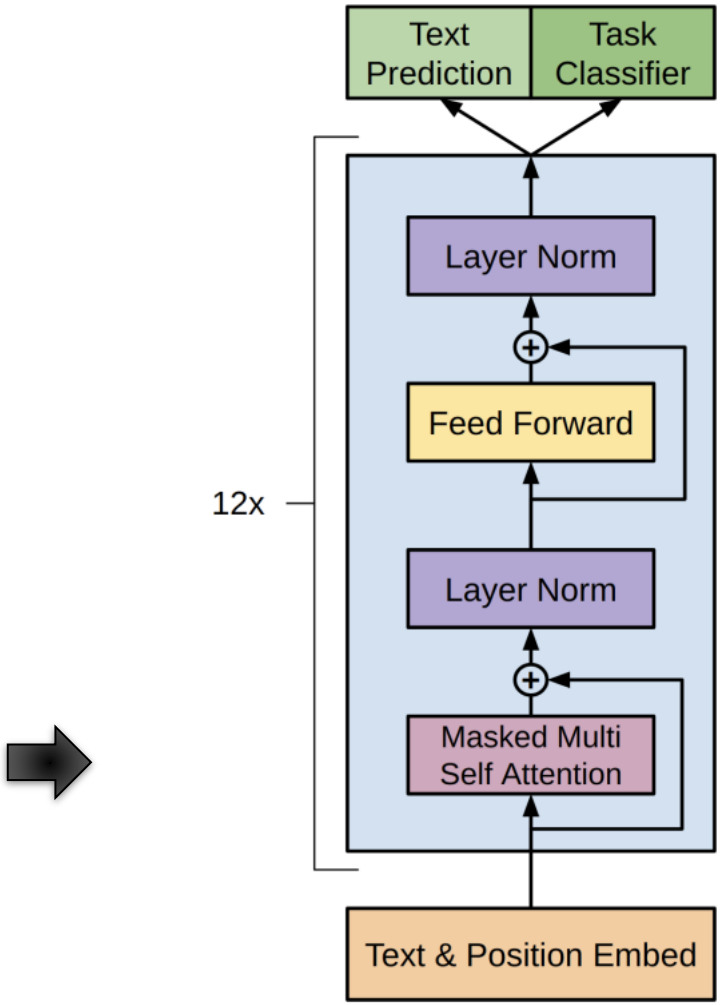
Mapping of each token ID to an embedding vector

Position Embeddings

Output

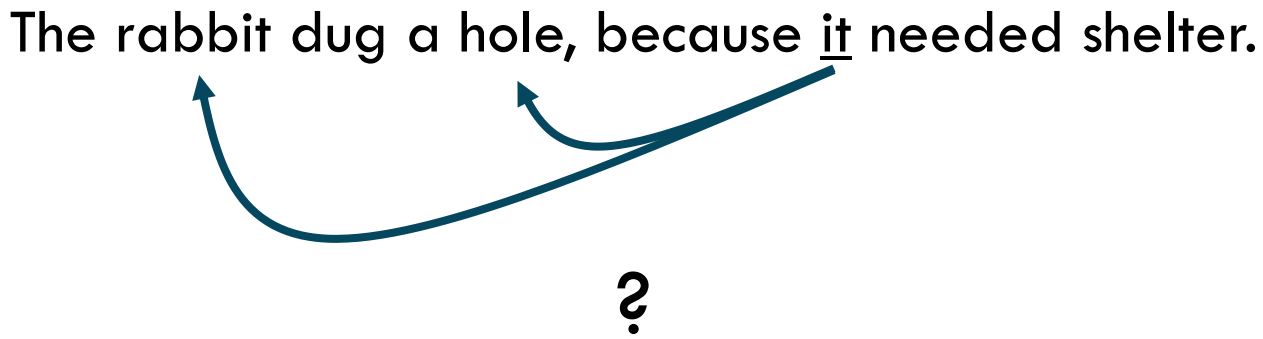


2) MODEL ARCHITECTURE



*3

MASKED MULTI SELF-ATTENTION



MASKED MULTI SELF-ATTENTION

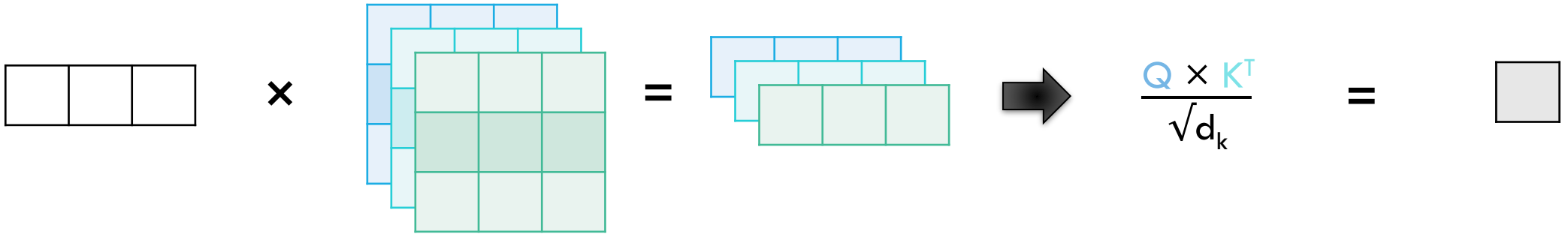
One Token Embedding Vector E

Weight matrices W_q, W_k and W_v $(d_{model} \times d_k)$

Query, key, and value matrices Q, K and V

score

score s



*$d_k = d_{model}$/ number of attention heads

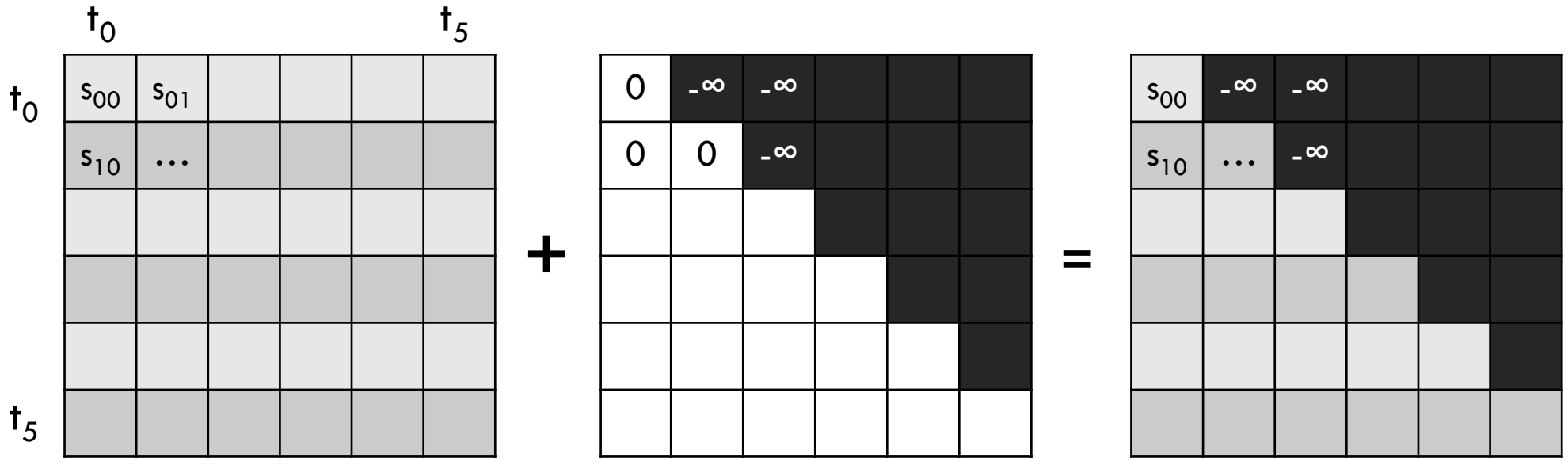
MASKED MULTI SELF-ATTENTION

For a sequence of 6 tokens:

Where each row represents the attention distribution for one specific token.

MASKED MULTI SELF-ATTENTION

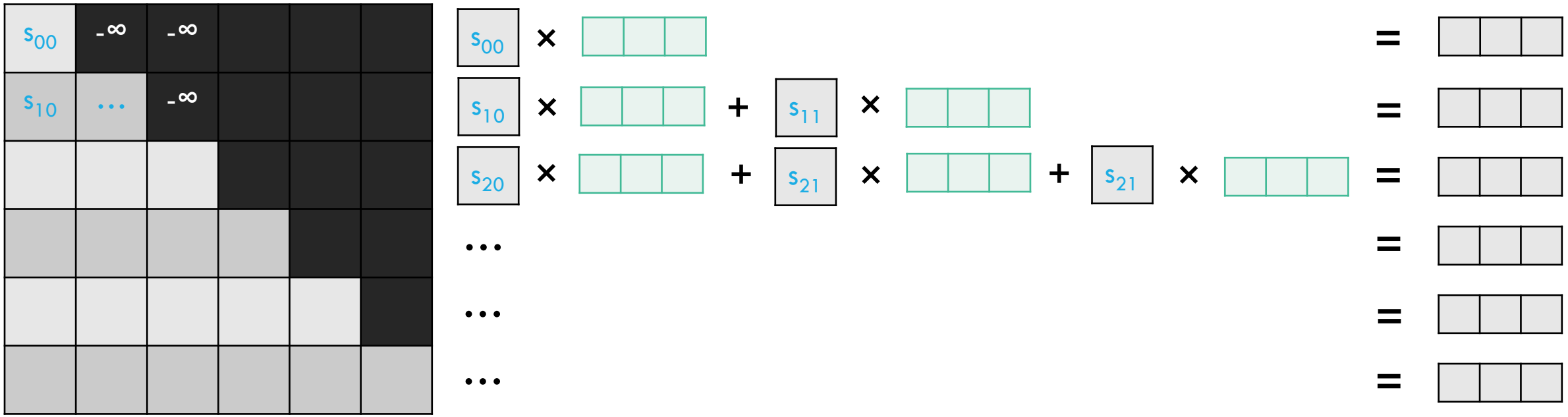
For a sequence of 6 tokens t_0 to t_5 with attention scores s_{00} to s_{55}.



*2

MASKED MULTI SELF-ATTENTION

Apply softmax-function to each score s and calculate $\sum_i(s_i \times V)$ for each row.

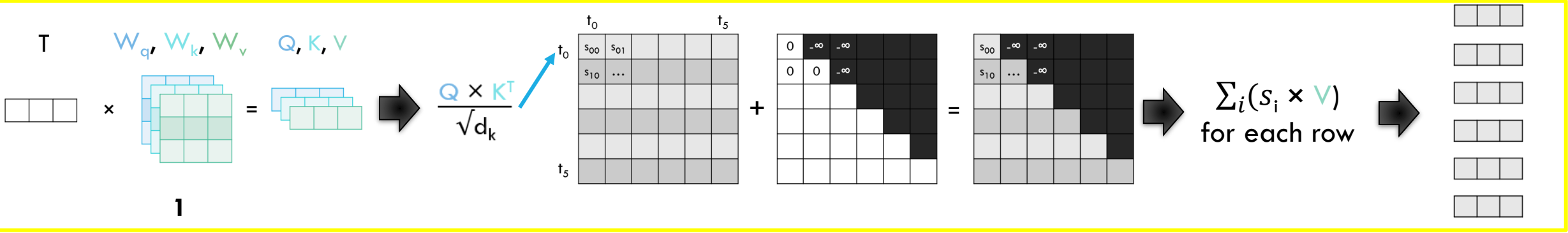


MASKED MULTI SELF-ATTENTION

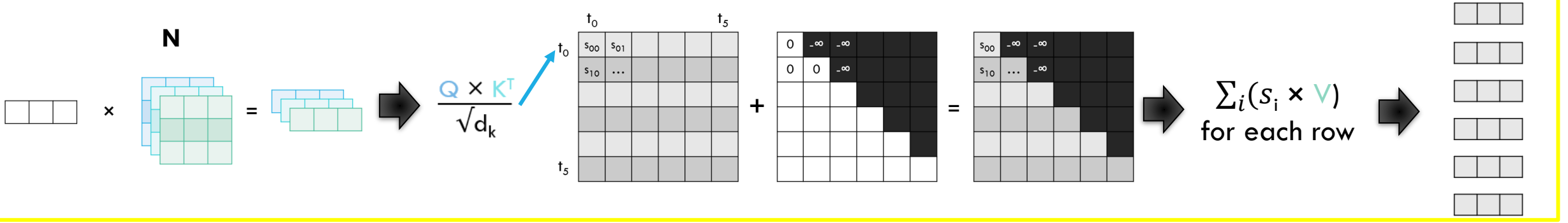
Compute score for each token

Masking on a sequence of tokens

Softmax & Attention score calculation



Attention Head

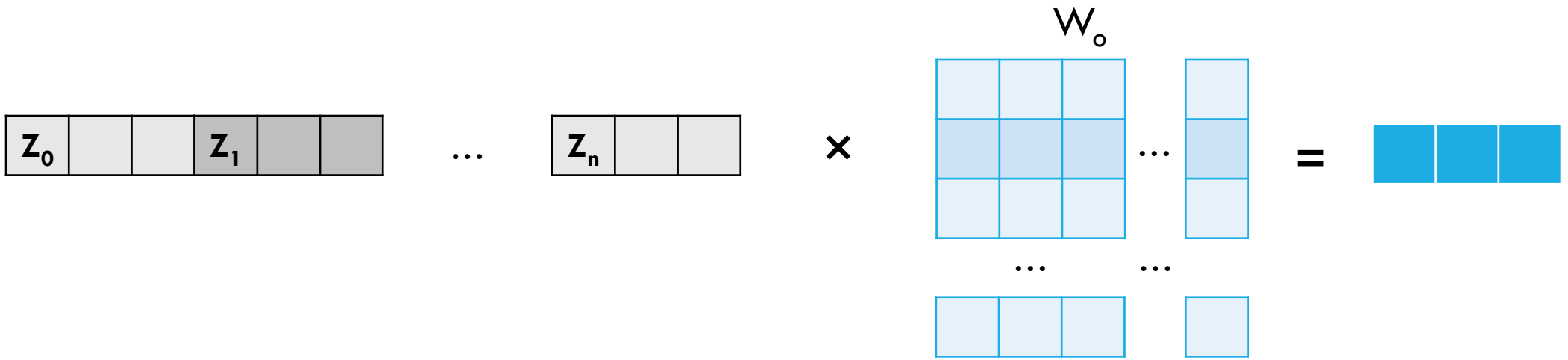


MASKED MULTI SELF-ATTENTION

Concatenate the output of all heads

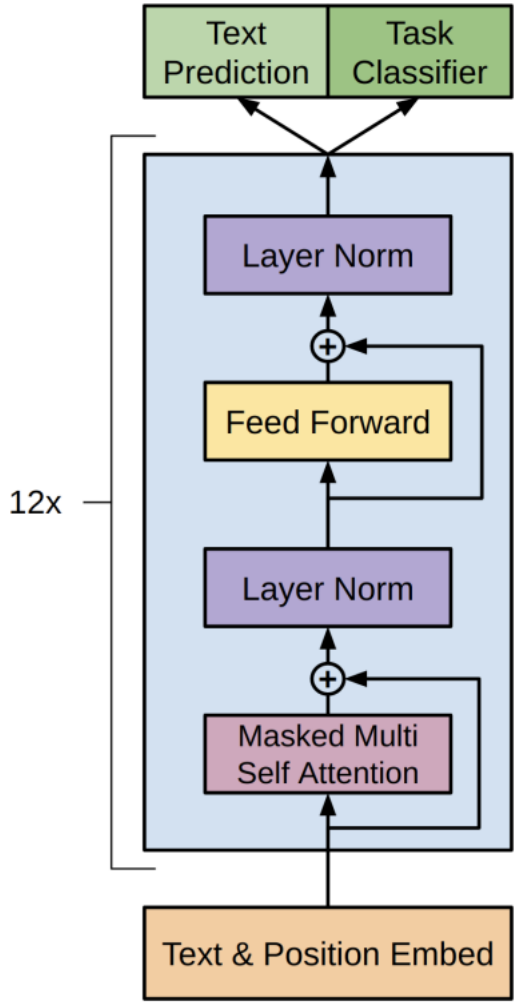
weight matrix

attention score



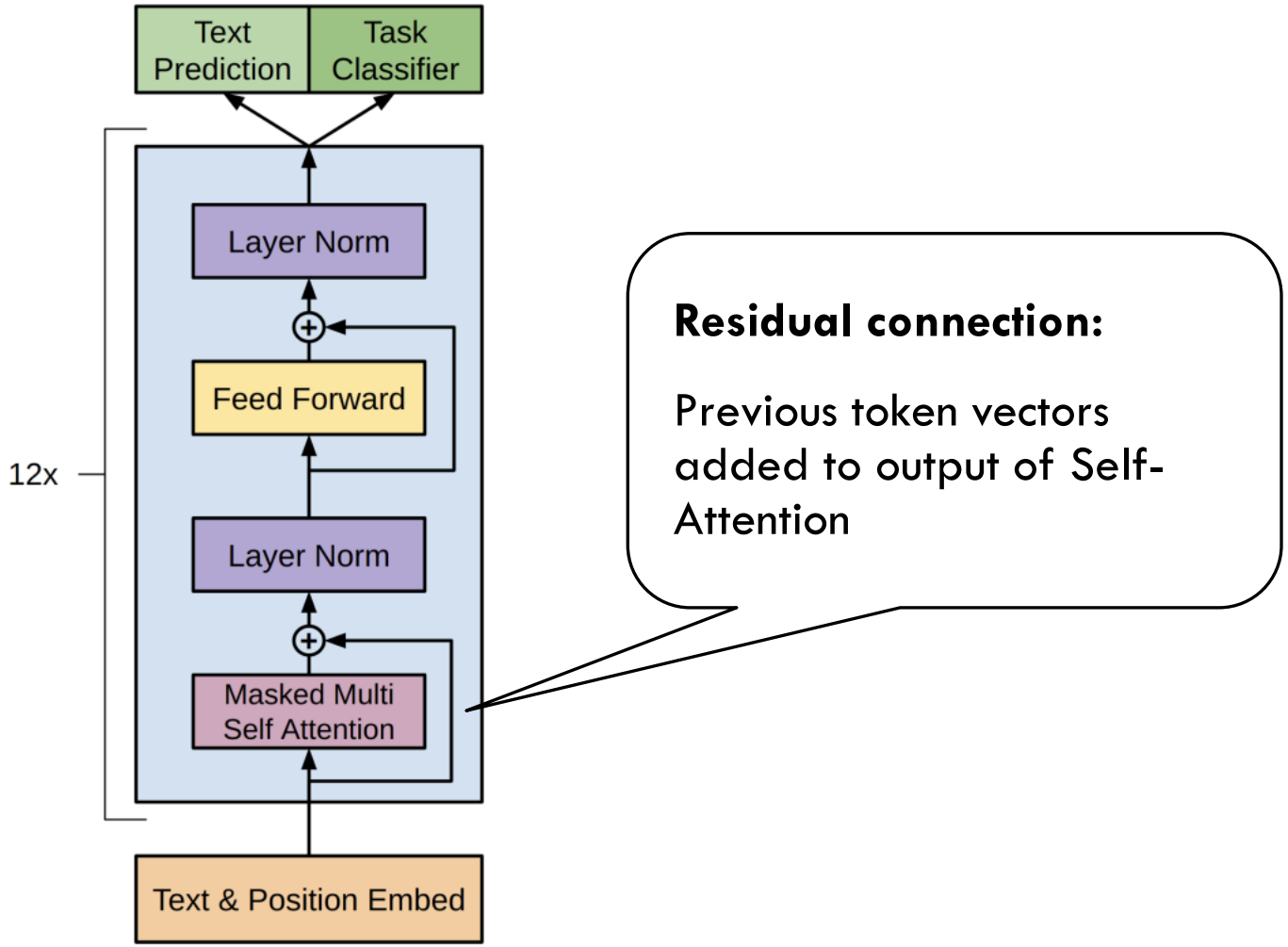
*1

2) MODEL ARCHITECTURE



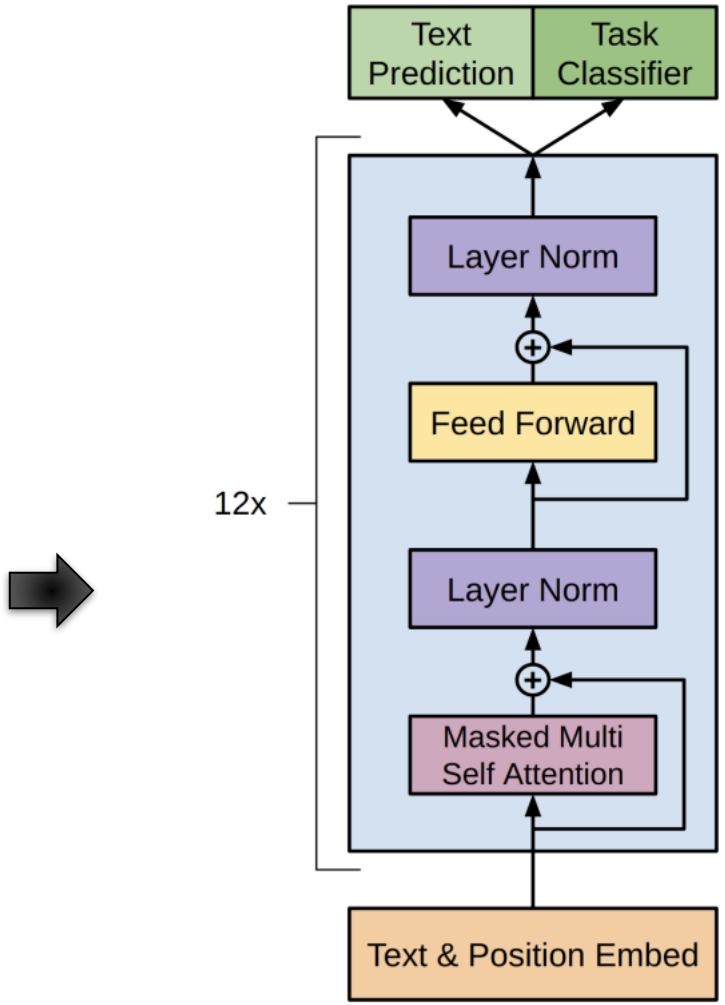
*3

2) MODEL ARCHITECTURE



*3

2) MODEL ARCHITECTURE



*3

LAYER NORM

Output of the Self-Attention layer:

x_0	x_1	...	x_n

Mean μ

$$\mu = \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Normalization equation y:

$$y = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \odot \gamma + \beta$$

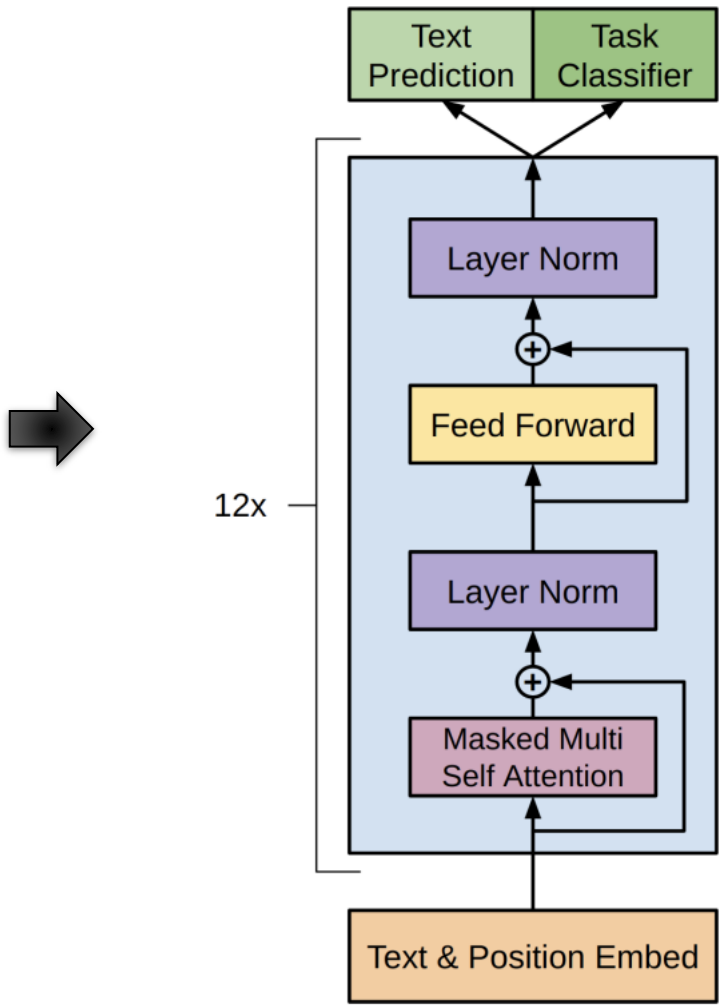
With trainable parameters γ and β

and small positve value ϵ

Standard deviation σ:

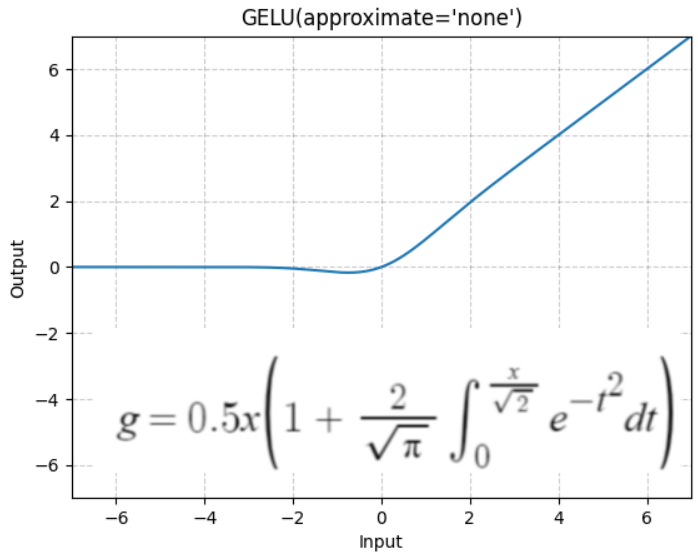
$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

2) MODEL ARCHITECTURE

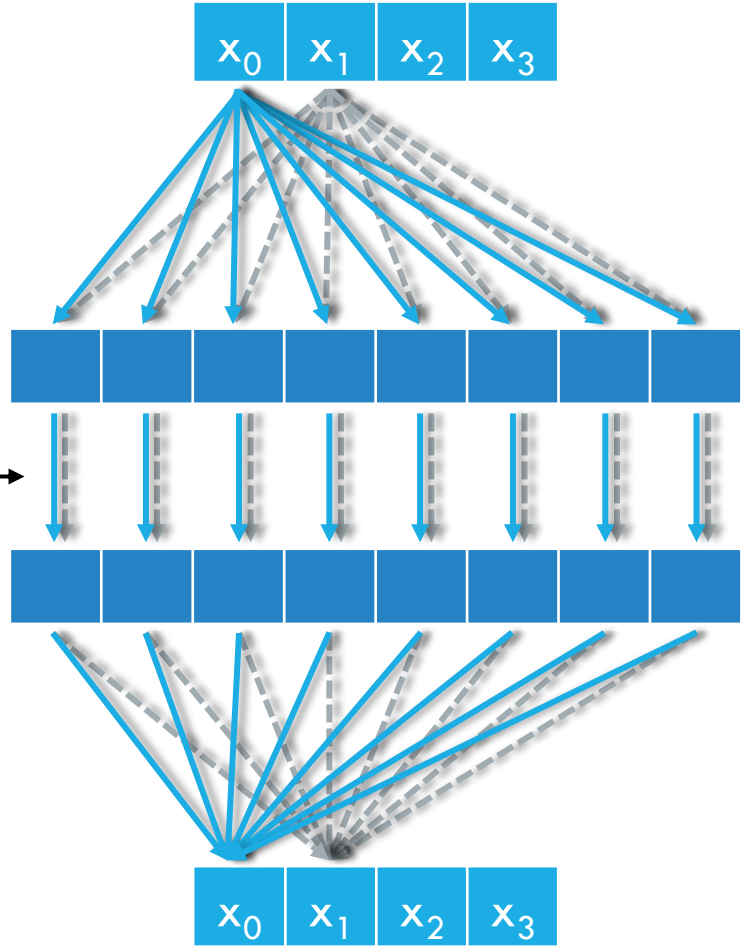


*3

FEED FORWARD LAYER

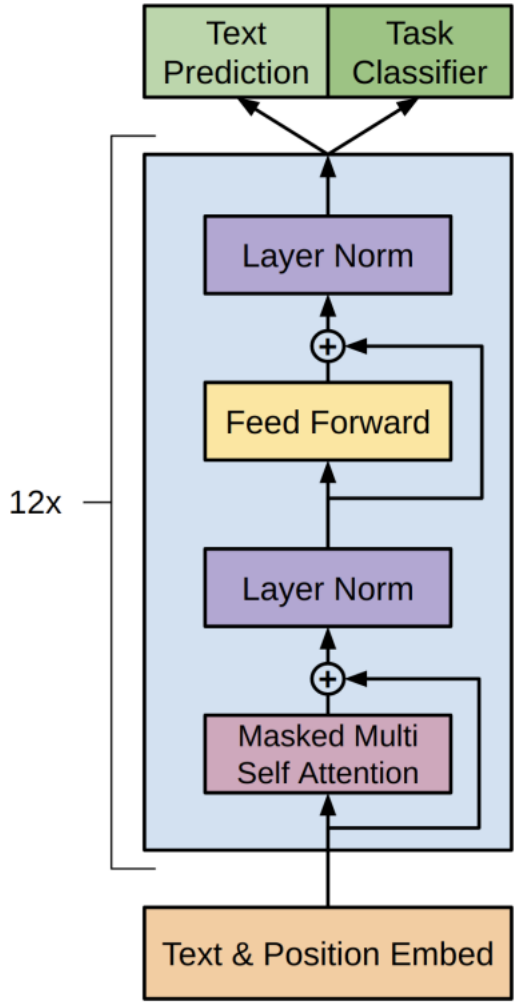


GELU-Activation



*5

2) MODEL ARCHITECTURE



*3

3) FRAMEWORK

1) Unsupervised pre-training
2) Supervised fine-tuning

3.1) UNSUPERVISED PRE-TRAINING

Training: 7k unpublished books

Goal: Learn a general language structure.

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

How does this sentence continue?

„The rabbit dug a …“

3.1) UNSUPERVISED PRE-TRAINING

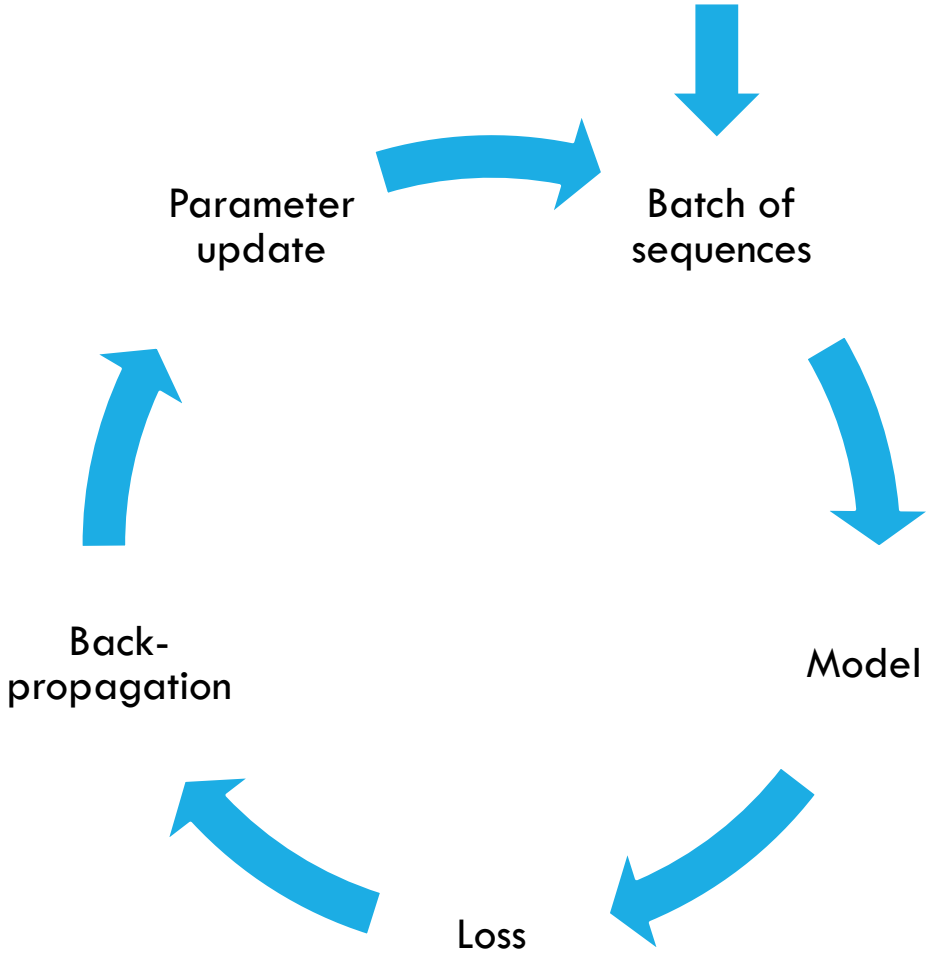
Goal: Learn a general language structure.

How does this sentence continue?

„The rabbit dug a hole“

```
Word:    hole        Probability:    0.4542
Word:    little      Probability:    0.0721
Word:    small       Probability:    0.0213
Word:    deep        Probability:    0.0197
Word:    few         Probability:    0.0197
Word:    path        Probability:    0.0151
Word:    grave       Probability:    0.0140
Word:    bit         Probability:    0.0134
Word:    trench      Probability:    0.0120
Word:    long        Probability:    0.0111
```

3.1) UNSUPERVISED PRE-TRAINING - TRAINING LOOP



3.2) SUPERVISED FINE-TUNING

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

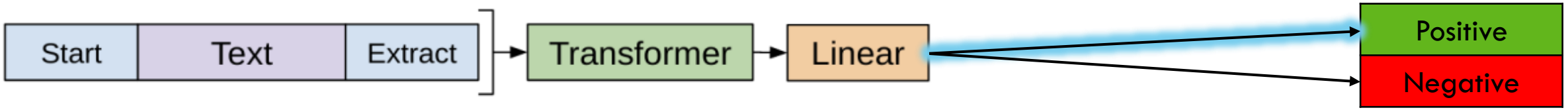
$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C})$$

From Pre-Training:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i|u_{i-k}, \ldots, u_{i-1}; \Theta)$$

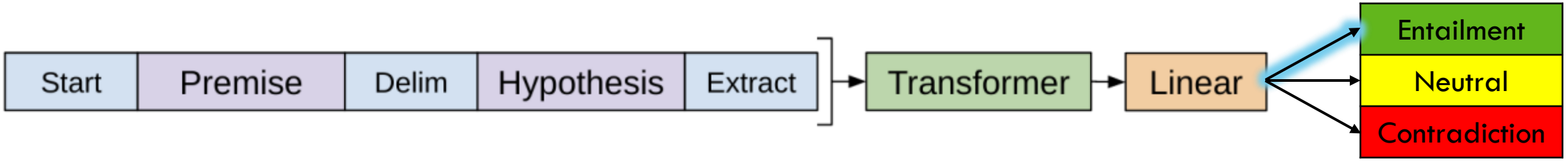
3.2) SUPERVISED FINE-TUNING

1) Classification: „The movie is fantastic!“



2) Entailment:

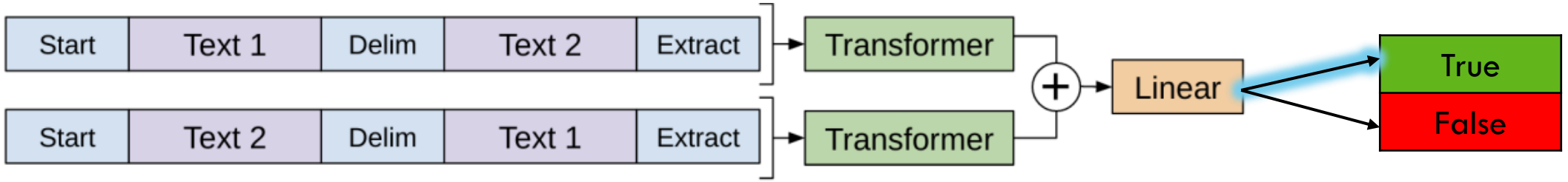
„The red cat sits on the high roof.“ „The cat sits on the roof.“



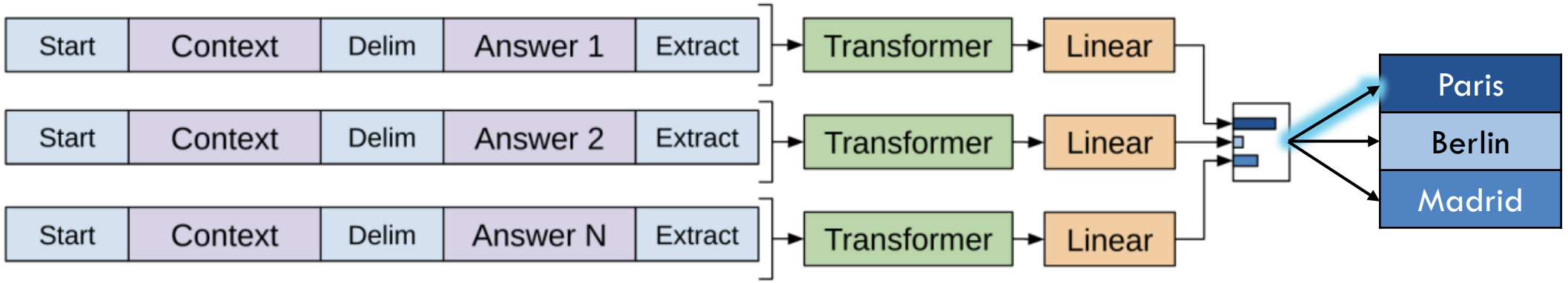
*3

3.2) SUPERVISED FINE-TUNING

3) Similarity: "She loves reading." „She enjoys reading."

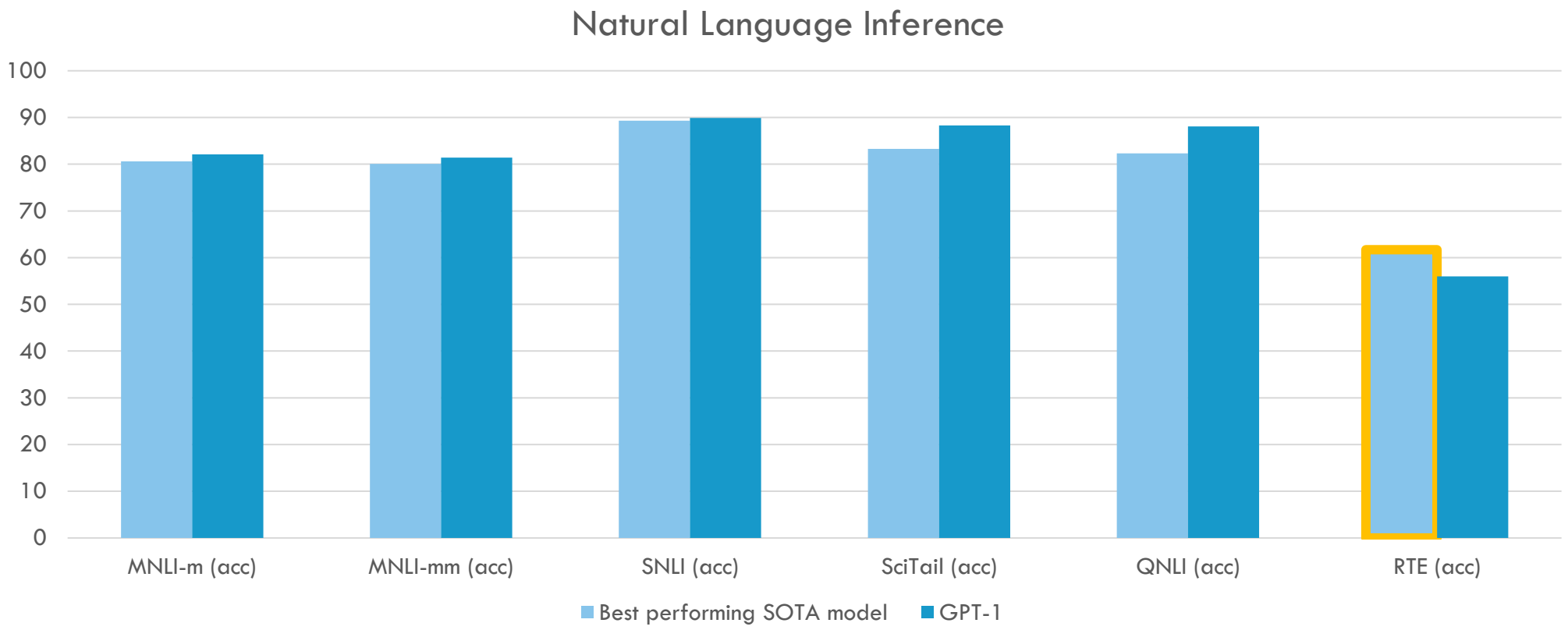


4) Multiple Choice: „What is the capital of France?"

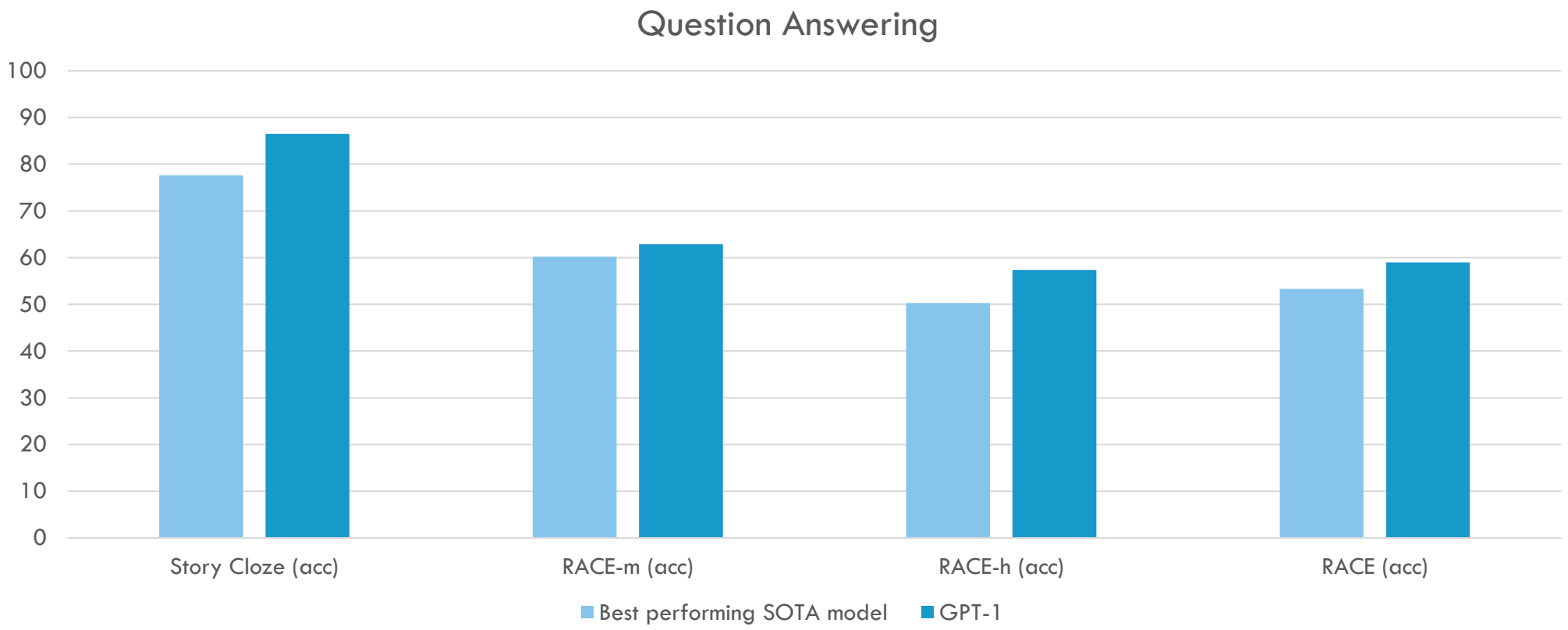


*3

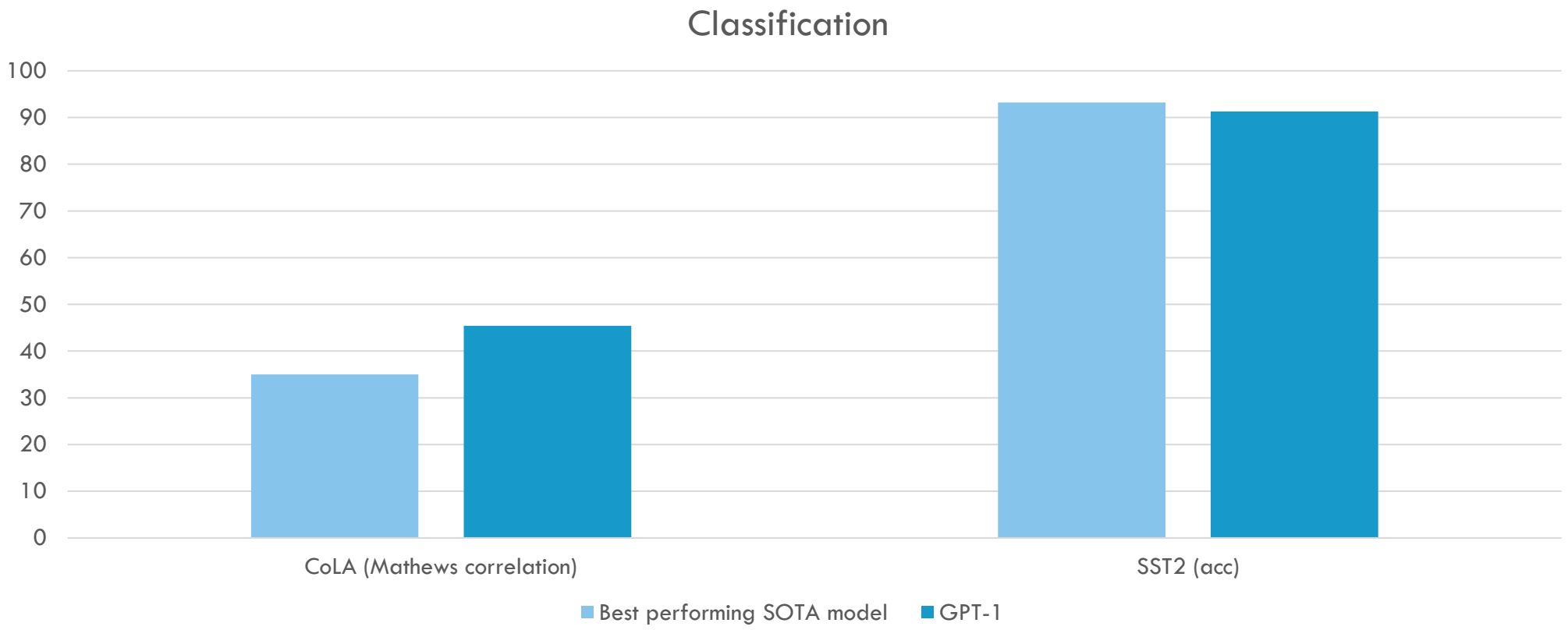
4) EVALUATION - NLI



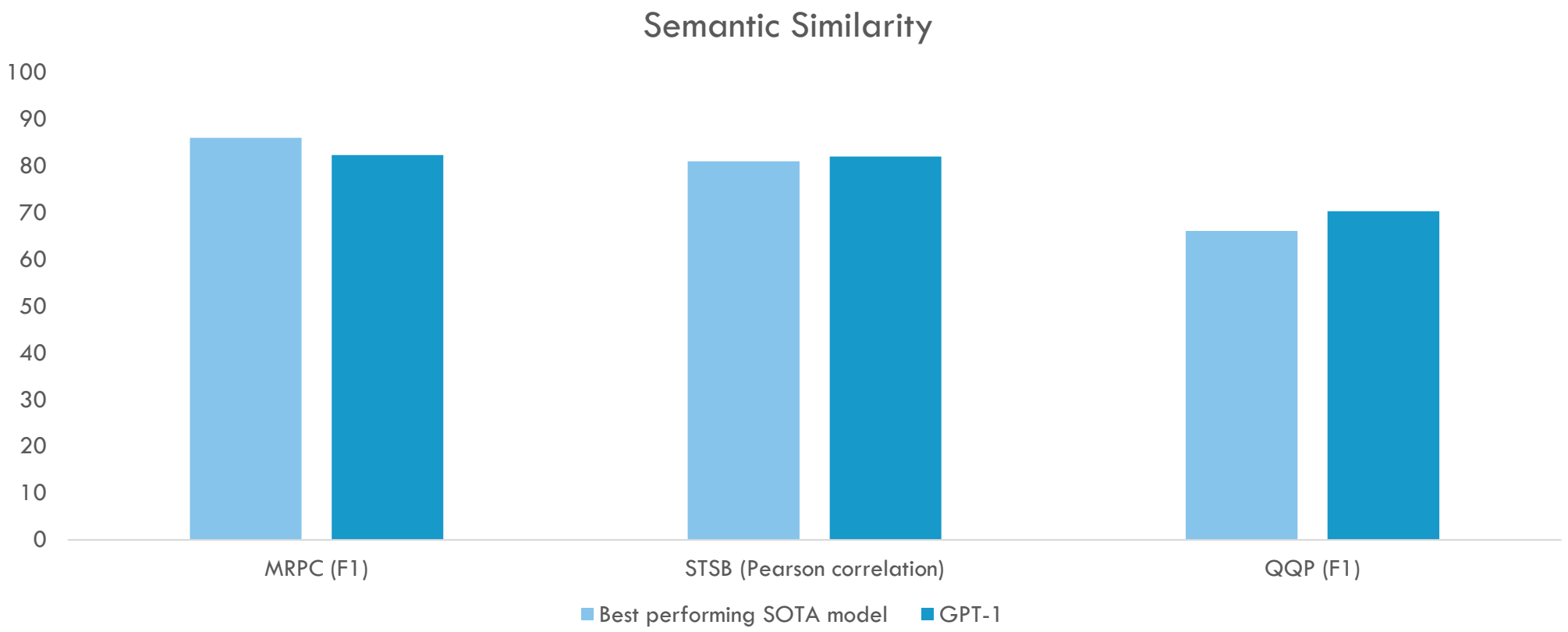
4) EVALUATION



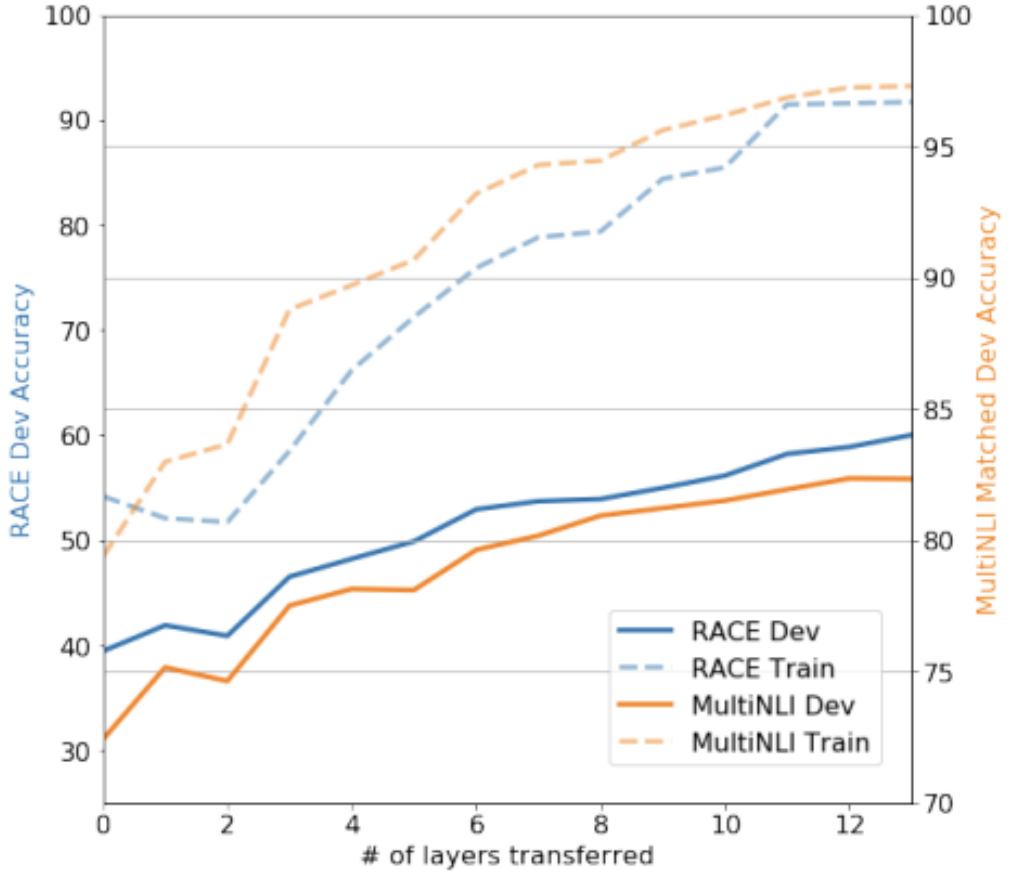
4) EVALUATION



4) EVALUATION

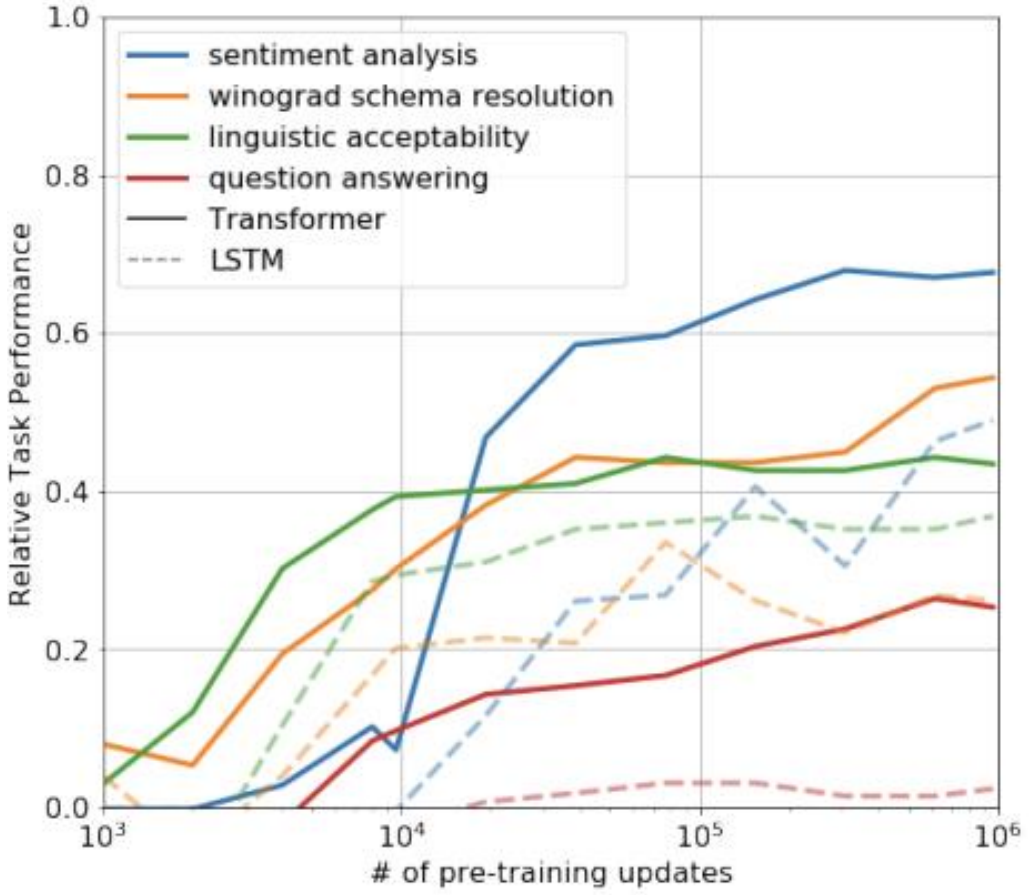


4) EVALUATION - LAYER TRANSFER



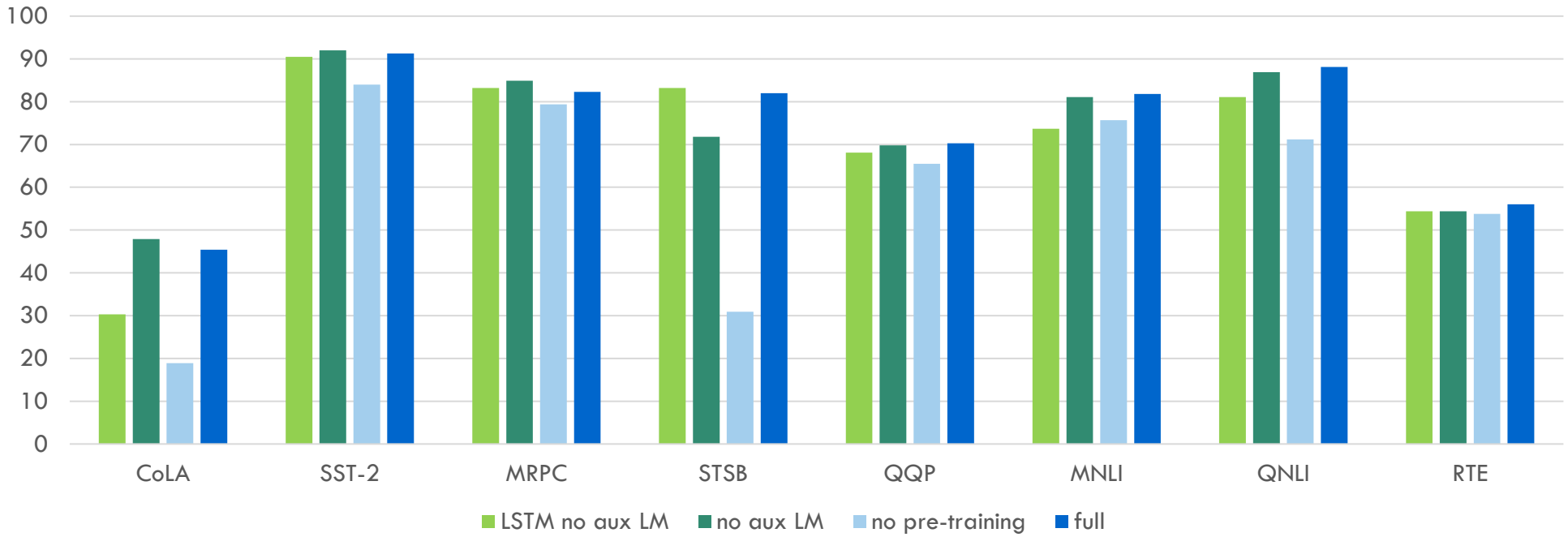
*3

4) EVALUATION - ZERO-SHOT BEHAVIORS

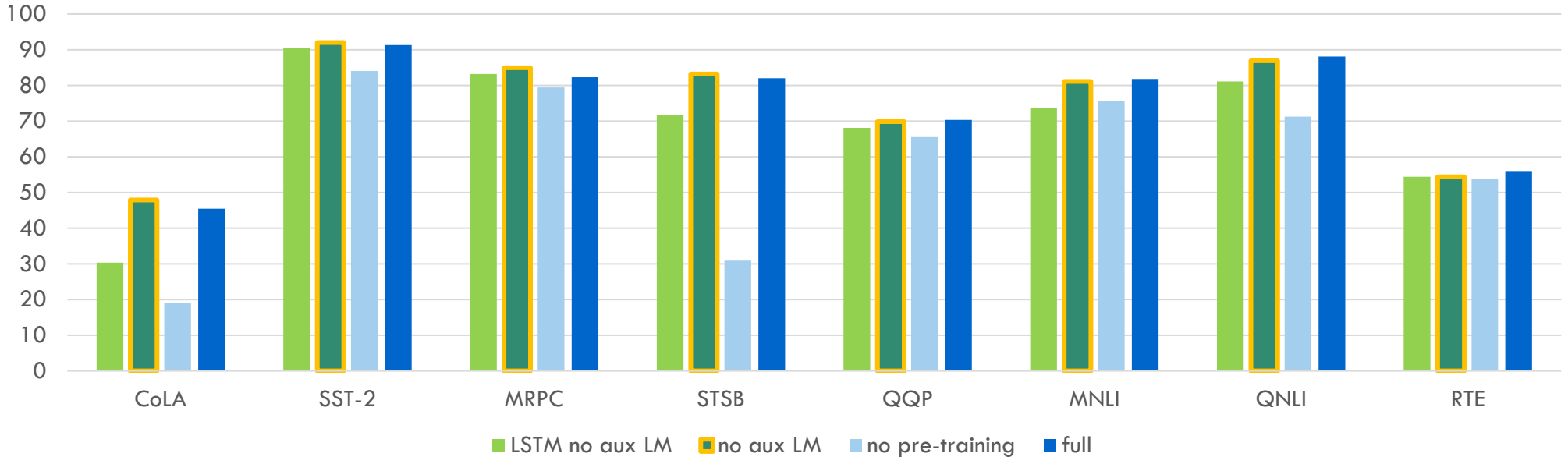


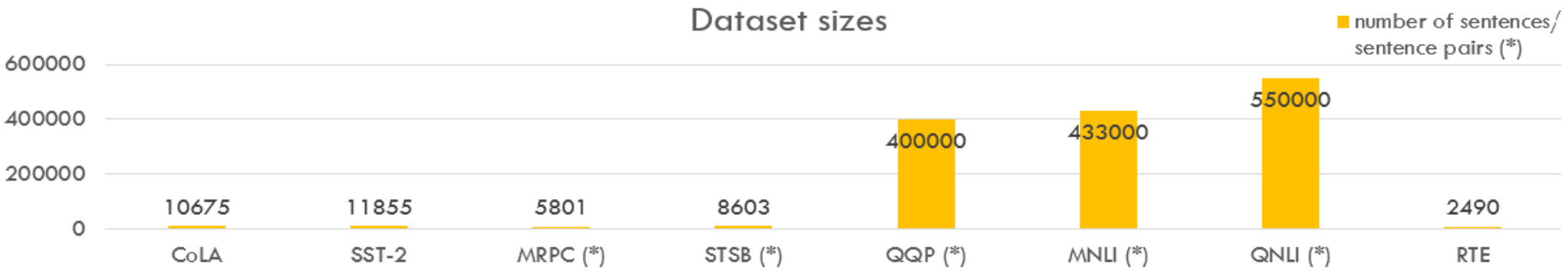
*3

4) EVALUATION - ABLATION STUDY



4) EVALUATION

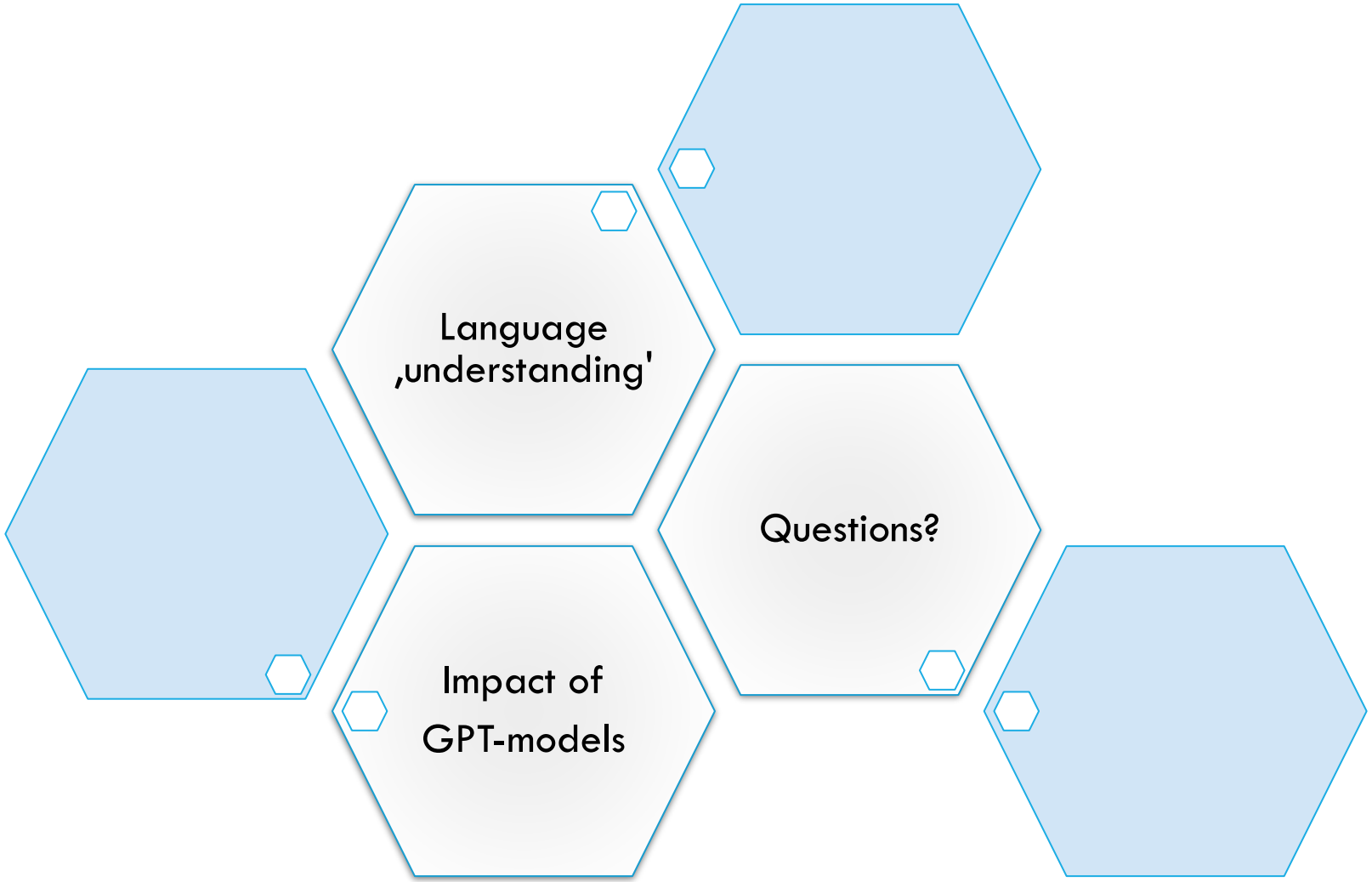




5) CONCLUSION

Captures long-range dependencies	Lack of domain specific knowledge
Task-agnostic	Limited effectiveness of supervised finetuning, particularly on smaller datasets
Unsupervised pre-training	The title: Improving language understanding by generative pre-training
Performance	

6) DISCUSSION



RECOMMENDATIONS

Intuitive transformer explanation:

https://jalammar.github.io/illustrated-transformer/

3Blue1Brown on Youtube (last two videos of the playlist):

https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

The paper:

https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

PICTURE SOURCES

*1 https://jalammar.github.io/illustrated-transformer/

*2 https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

*3

https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

*4 https://medium.com/@hunter-j-phillips/layer-normalization-e9ae93eb3c9c

*5 https://datascience.stackexchange.com/questions/49522/what-is-gelu-activation

Thank you for your attention!