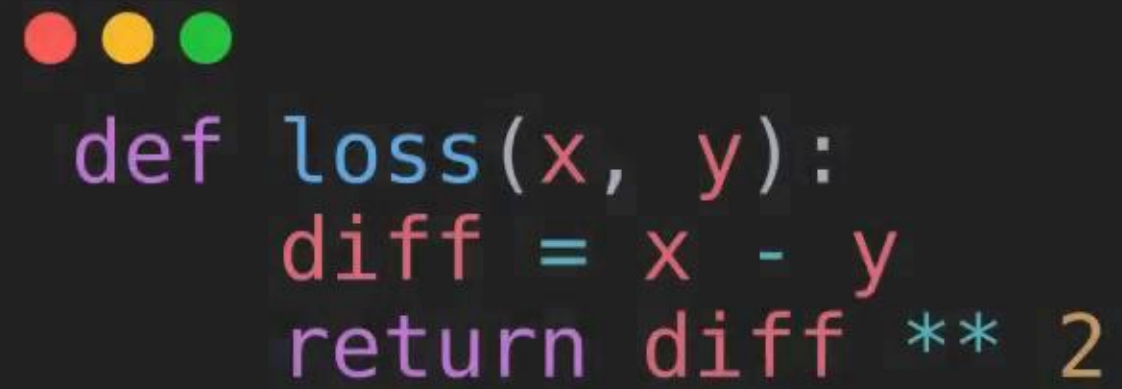


Names Don't Matter: Symbol-Invariant Transformer for Open-Vocabulary Learning

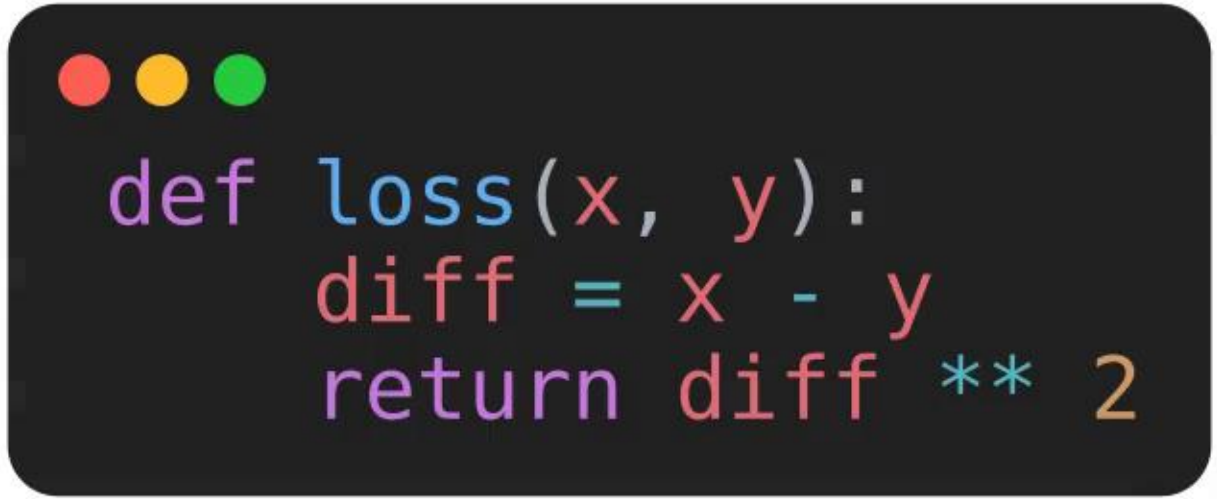
İlker Işık, Wenchao Li

Variable Renaming Problem



```
def loss(x, y):  
    diff = x - y  
    return diff ** 2
```

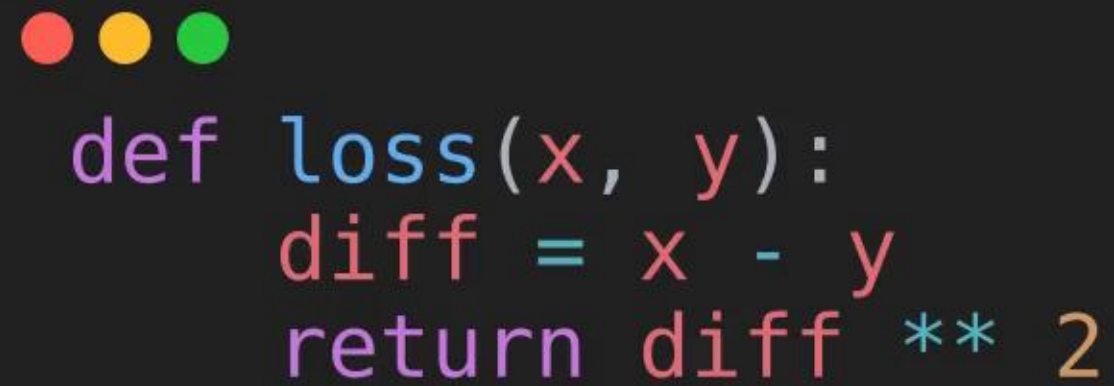
Variable Renaming Problem

A dark-themed code editor window with three colored window control buttons (red, yellow, green) in the top-left corner. It contains a Python function definition for a squared loss function.

```
def loss(x, y):  
    diff = x - y  
    return diff ** 2
```

A standard squared loss function.

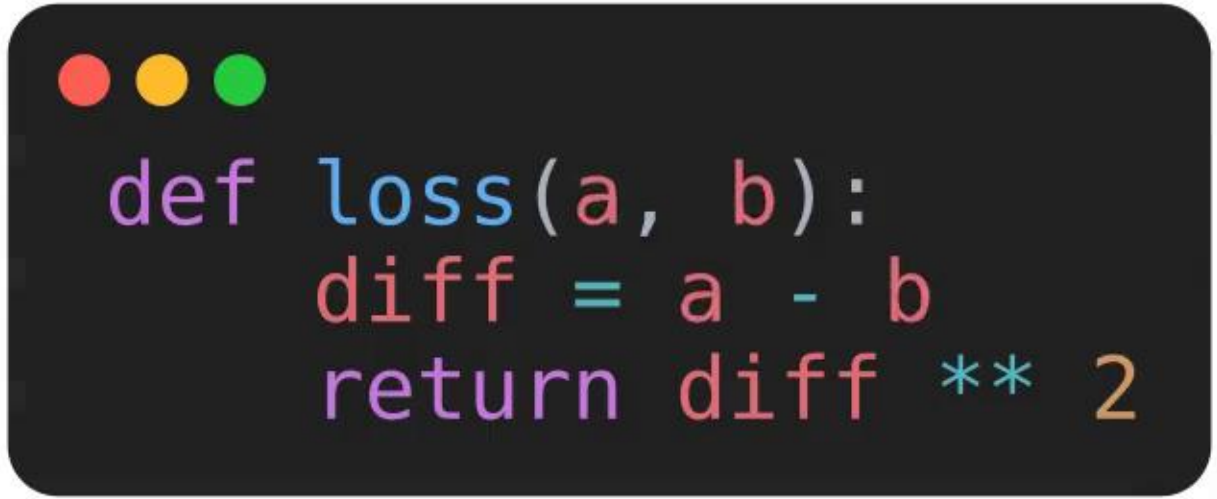
Variable Renaming Problem



```
def loss(x, y):  
    diff = x - y  
    return diff ** 2
```

Rename the variables ...

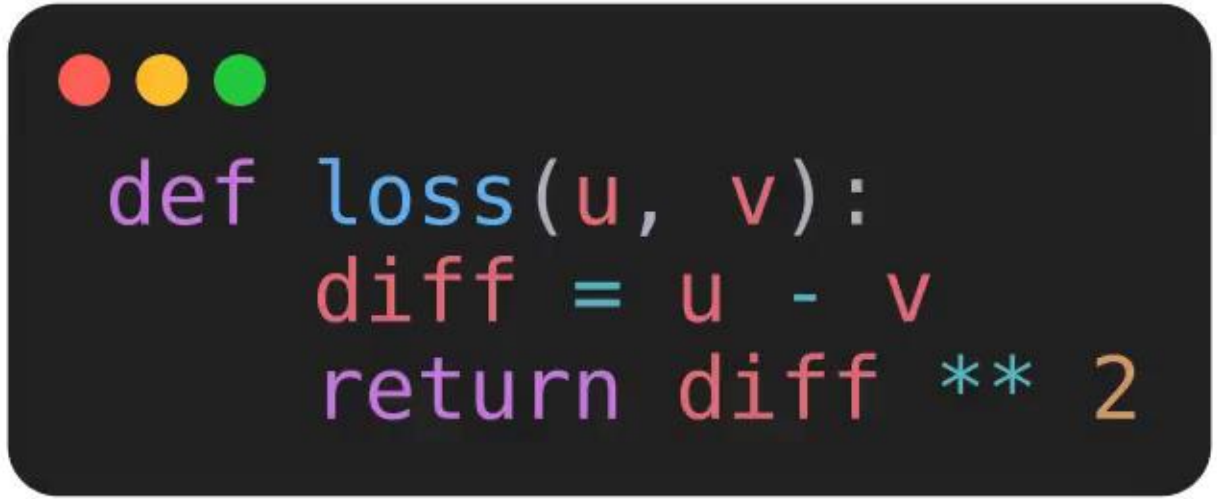
Variable Renaming Problem

A dark-themed code editor window with three colored window control buttons (red, yellow, green) in the top-left corner. It contains a Python function definition for a loss function.

```
def loss(a, b):  
    diff = a - b  
    return diff ** 2
```

Rename the variables ...

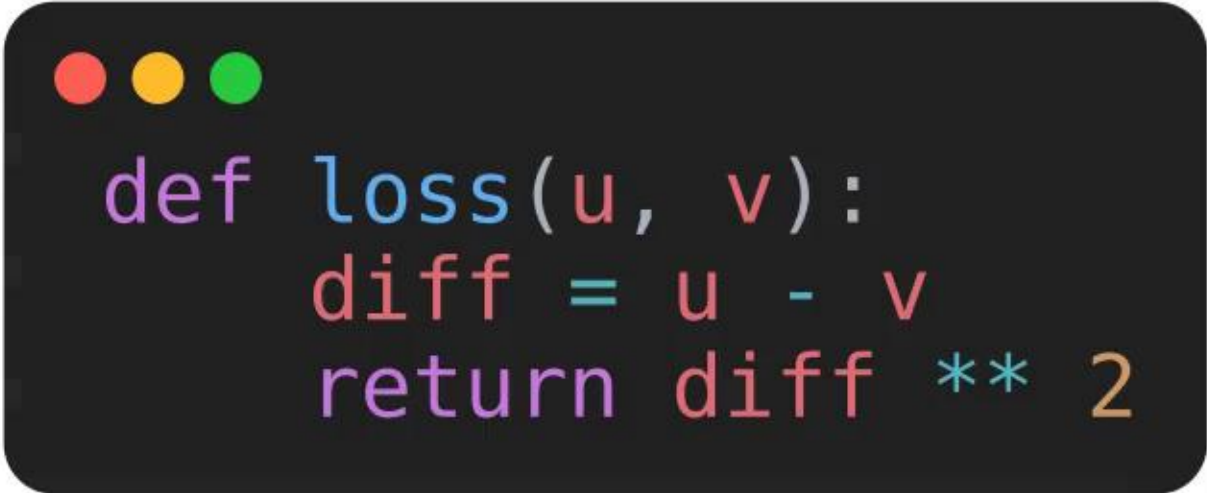
Variable Renaming Problem



```
def loss(u, v):  
    diff = u - v  
    return diff ** 2
```

Rename the variables ...

Variable Renaming Problem



```
def loss(u, v):  
    diff = u - v  
    return diff ** 2
```

Same computation — regardless of variable names.
Do ML models know this?

Impact of Renaming



```
def loss(pred, target):  
    diff = pred - target  
    return diff ** 2
```

High Accuracy

Impact of Renaming



```
def loss(pred, target):  
    diff = pred - target  
    return diff ** 2
```

rename
→



```
def loss(x, y):  
    diff = x - y  
    return diff ** 2
```

High Accuracy

Impact of Renaming



```
def loss(pred, target):  
    diff = pred - target  
    return diff ** 2
```

rename
→



```
def loss(x, y):  
    diff = x - y  
    return diff ** 2
```

Low Accuracy

Variable Renaming in Language

How much does Alice owe Bob?

Variable Renaming in Language

How much does Alice owe Bob?

Renaming the people doesn't change the question.

Variable Renaming in Language

How much does Alice owe Bob?

Renaming the people doesn't change the question.

Fixed Embeddings

Fixed Embeddings

Training Variables

a

b

c

Fixed Embeddings

Training Variables

a

b

c

Test Variables

a

b

c



Open-Vocabulary

Training Variables

a

b

c

Test Variables

a

b

c



Open-Vocabulary

Training Variables

a

b

c

Test Variables

a

b

c

d

e

f

$\vdots$



Standard Transformer

Standard Transformer

Input

$$a \wedge b \vee c$$

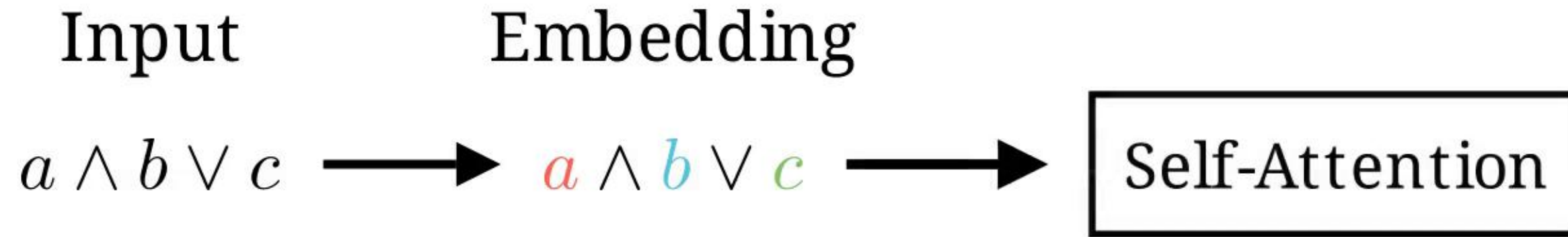
Standard Transformer

Input

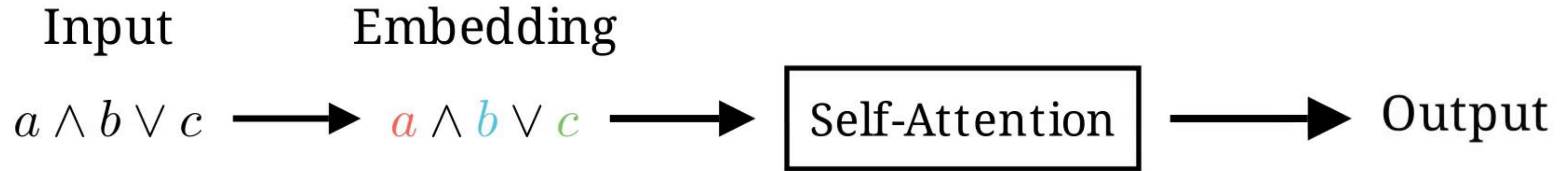
Embedding

$a \wedge b \vee c \longrightarrow \textcolor{red}{a} \wedge \textcolor{blue}{b} \vee \textcolor{green}{c}$

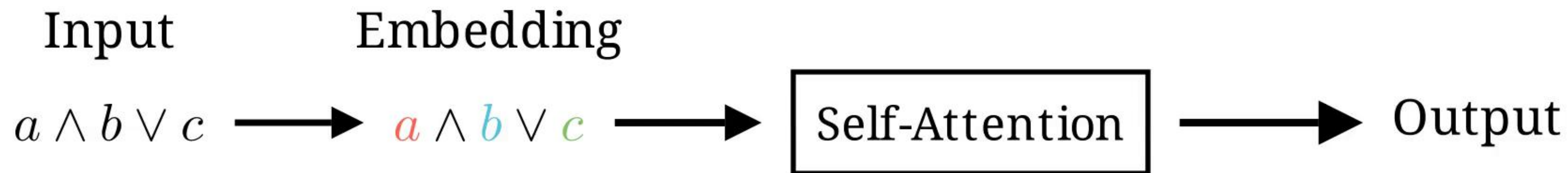
Standard Transformer



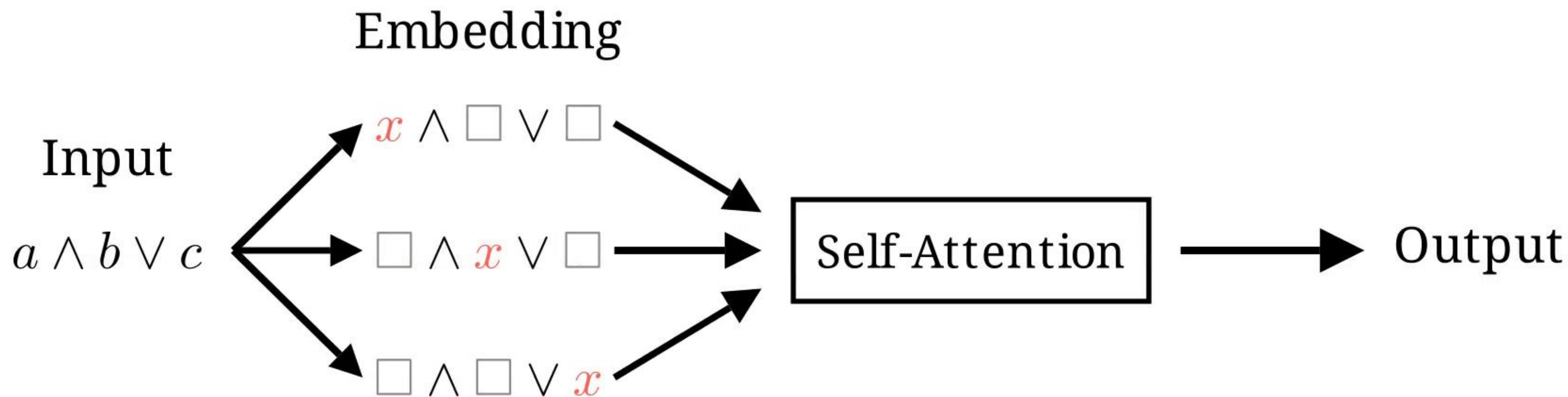
Standard Transformer



Symbol-Invariant Transformer

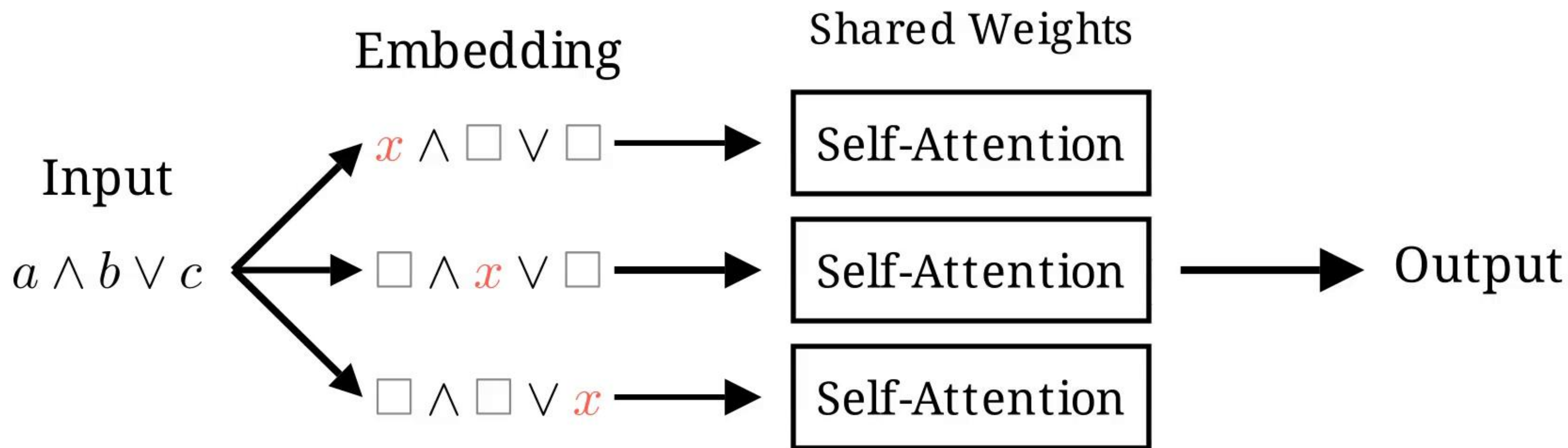


Symbol-Invariant Transformer



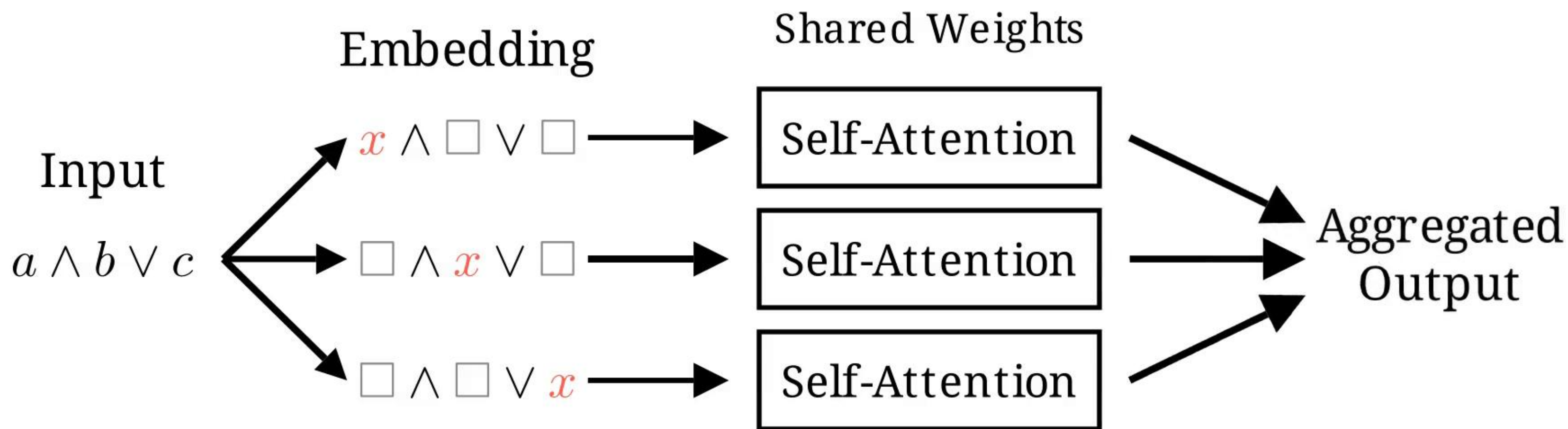
Create parallel embedding streams, one for each variable

Symbol-Invariant Transformer



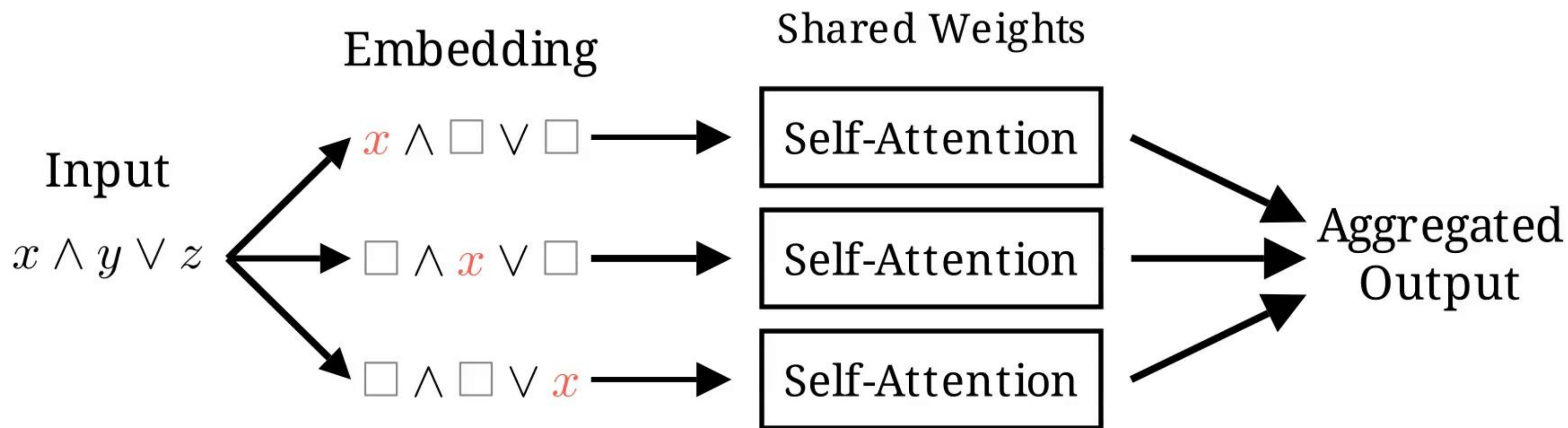
Process each stream independently with the same weights

Symbol-Invariant Transformer



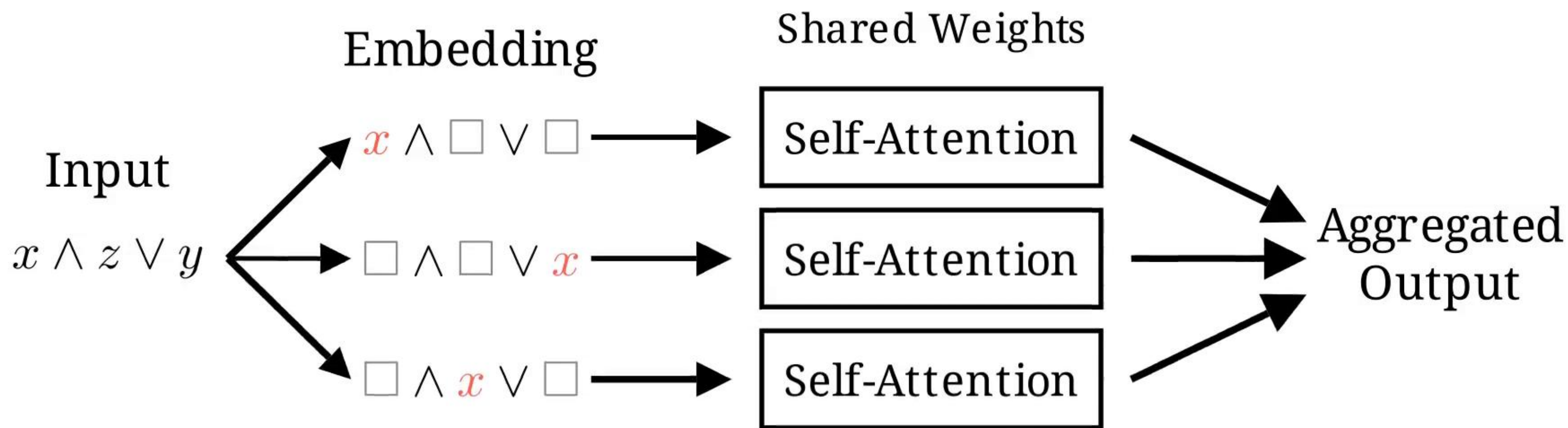
Aggregate output with a commutative operation

Symbol-Invariant Transformer



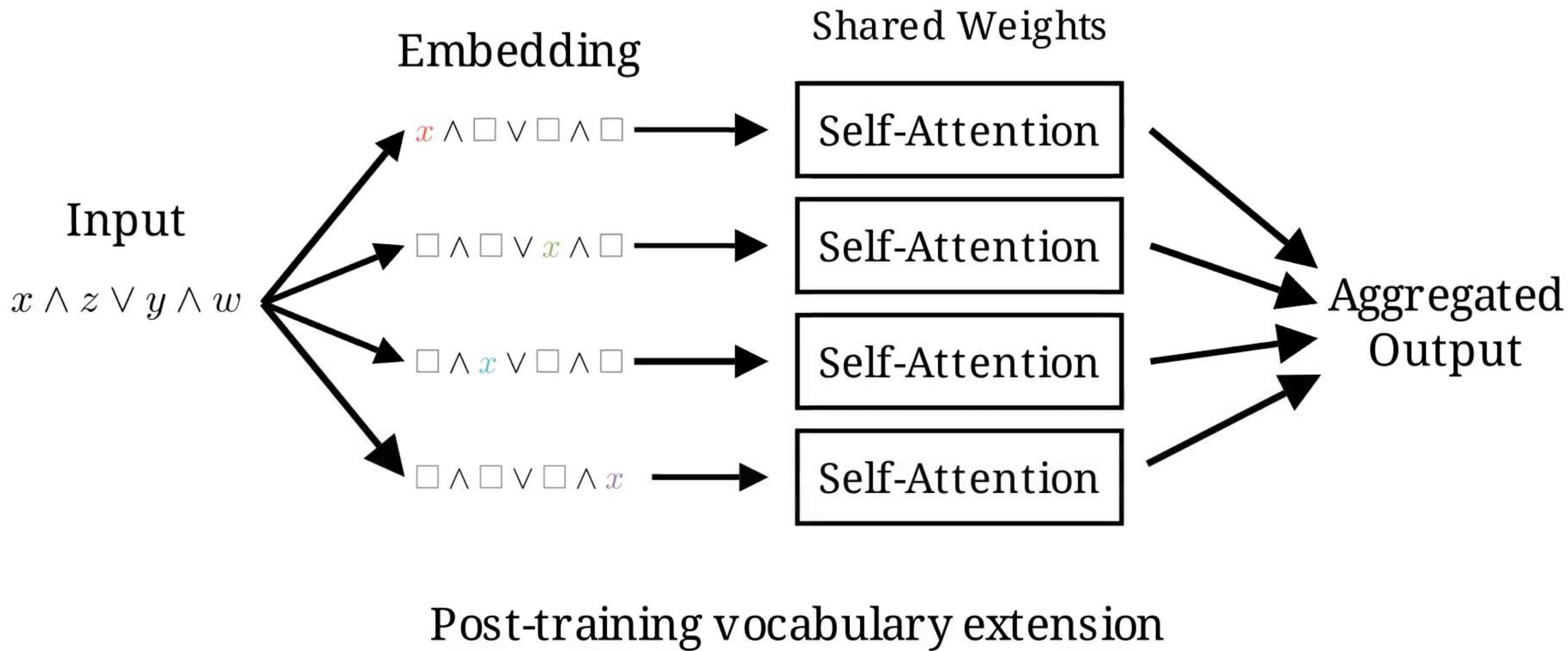
Renaming does not affect the embedding

Symbol-Invariant Transformer

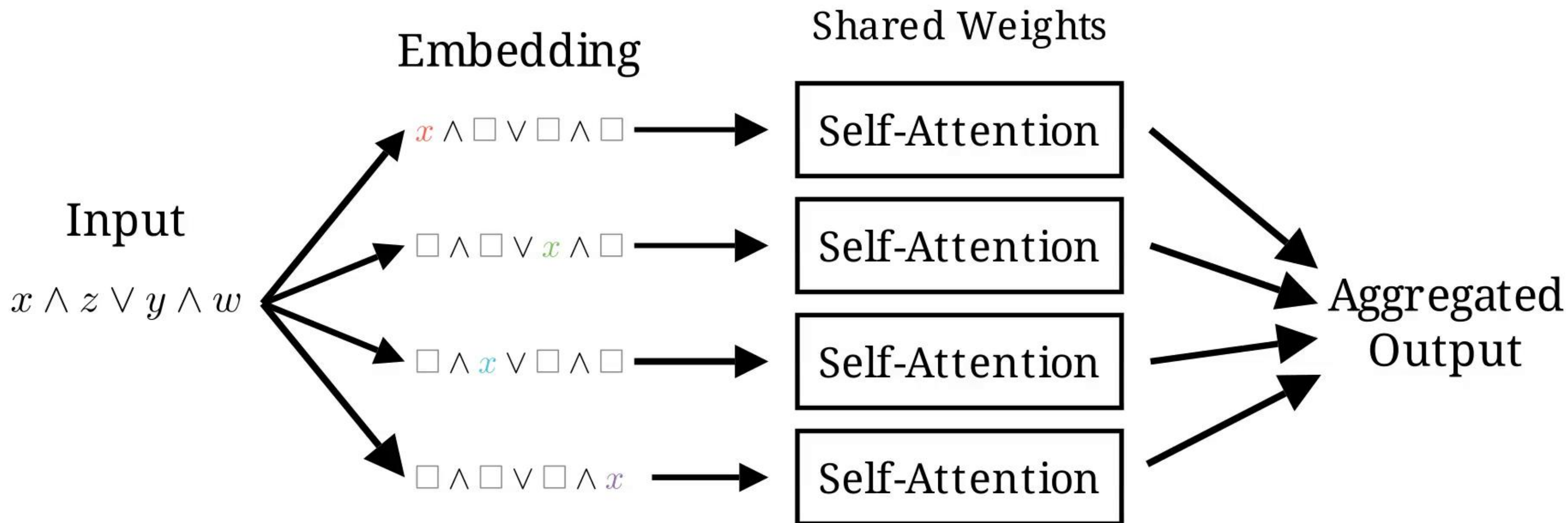


Stream order does not affect the embedding

Symbol-Invariant Transformer



Symbol-Invariant Transformer



See the full paper for other details

Experiments

Experiments

Empirical validation of renaming invariance

Experiments

Empirical validation of renaming invariance

Robustness against: training bias, reduced dataset

Experiments

Empirical validation of renaming invariance

Robustness against: training bias, reduced dataset

State-of-the-art generalization on logic tasks

Experiments

Empirical validation of renaming invariance

Robustness against: training bias, reduced dataset

State-of-the-art generalization on logic tasks

Fine-tune vanilla transformers into the proposed arch.

