Atten-Scope:

A Tool for Interpreting Large Language Model

Dongjae Lee

2024.05.30

KAIST
Programming Systems Laboratory

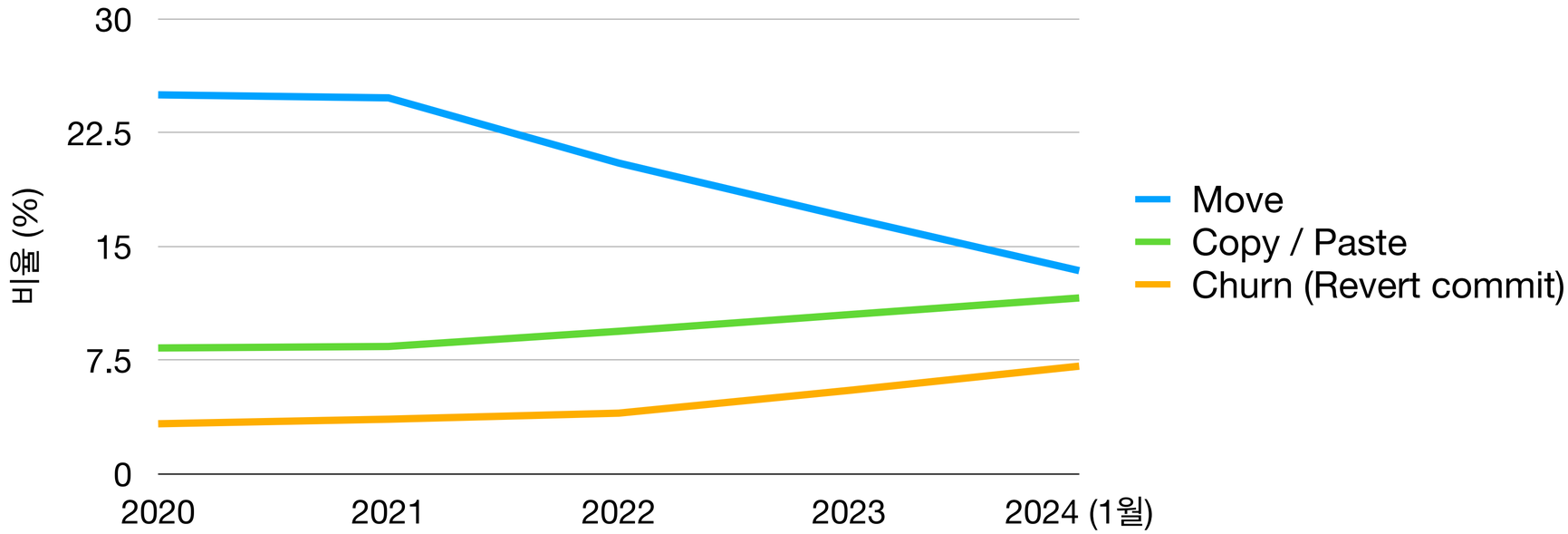
Importance of interpreting LLM

- Code quality has been declining since the language model came out*.

*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

Importance of interpreting LLM

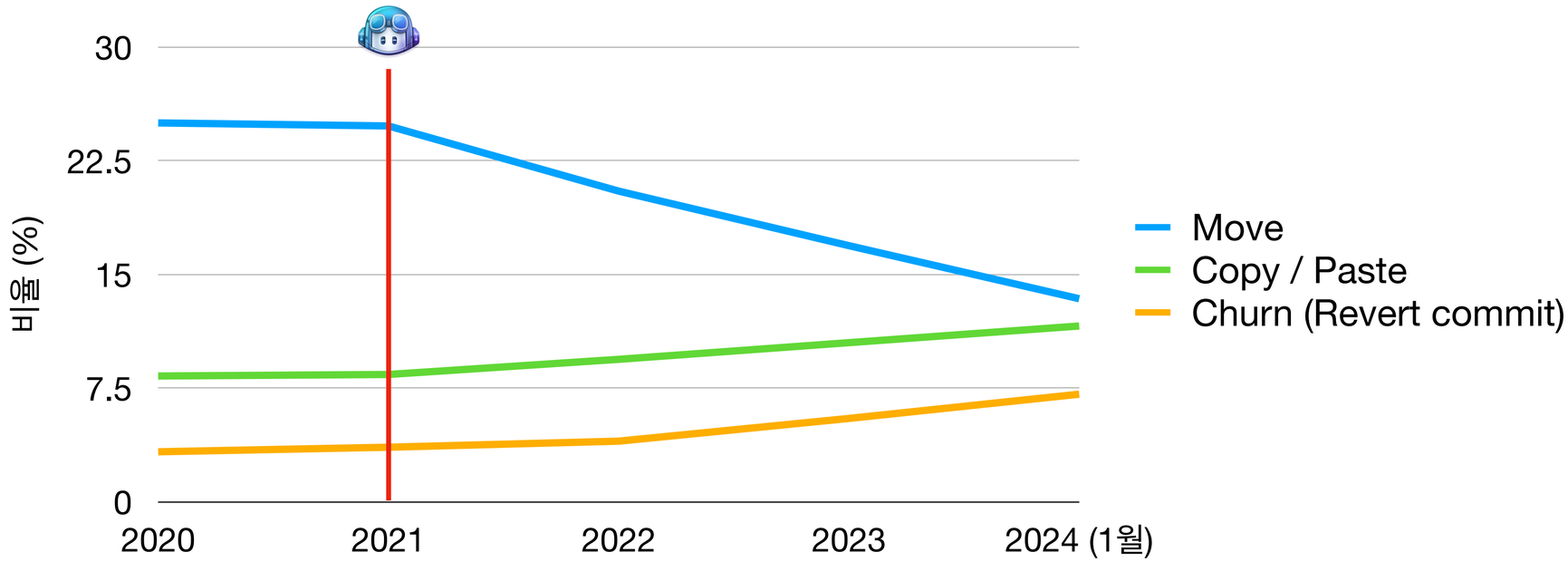
- Code quality has been declining since the language model came out*.



*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

Importance of interpreting LLM

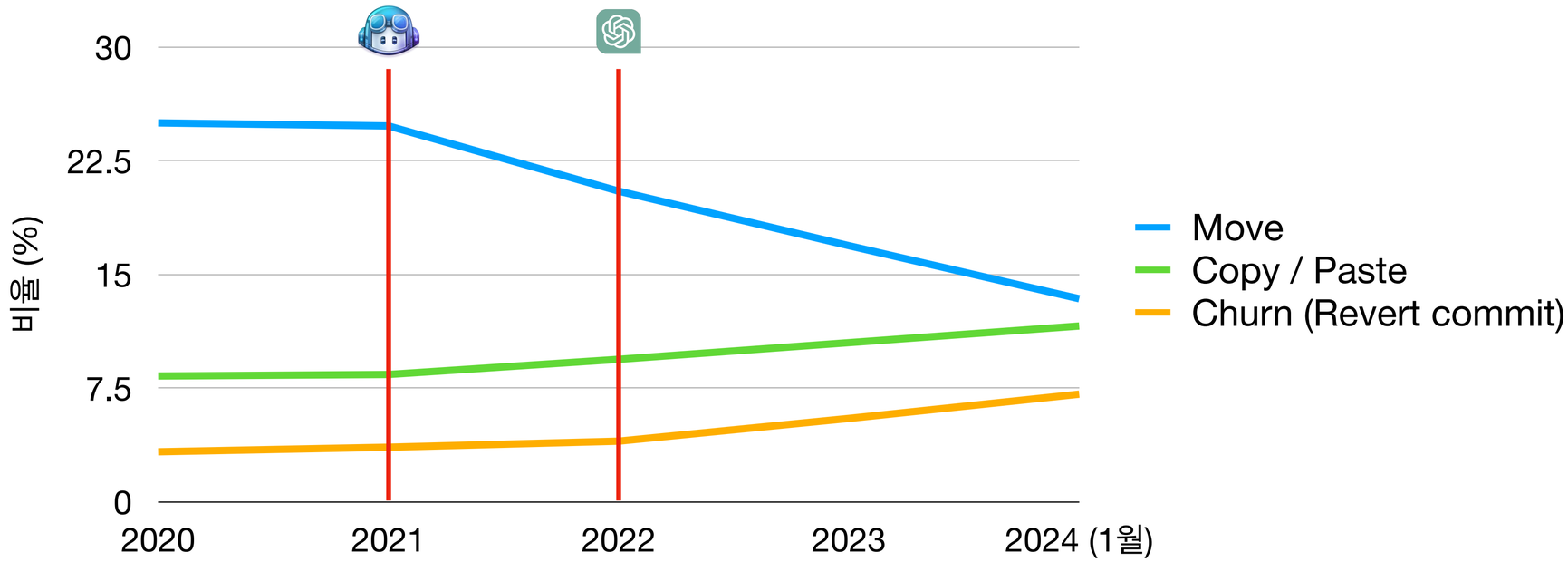
- Code quality has been declining since the language model came out*.



*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

Importance of interpreting LLM

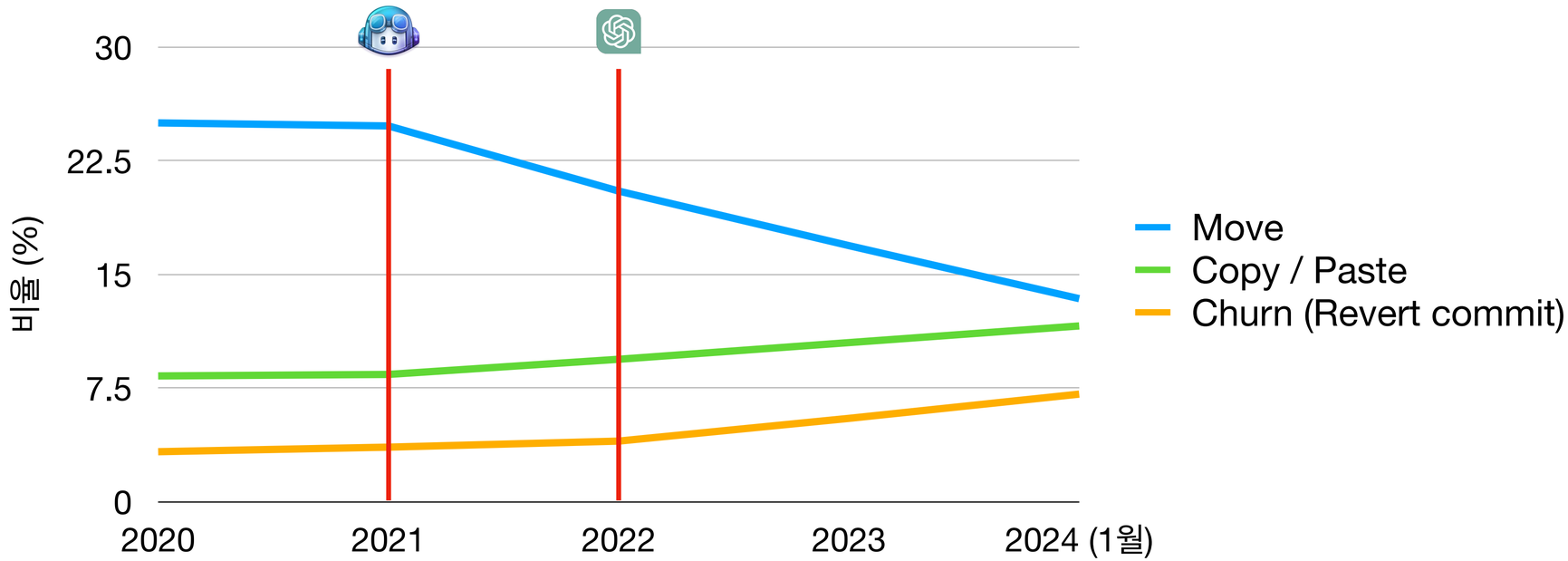
- Code quality has been declining since the language model came out*.



*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

Importance of interpreting LLM

- Code quality has been declining since the language model came out*.

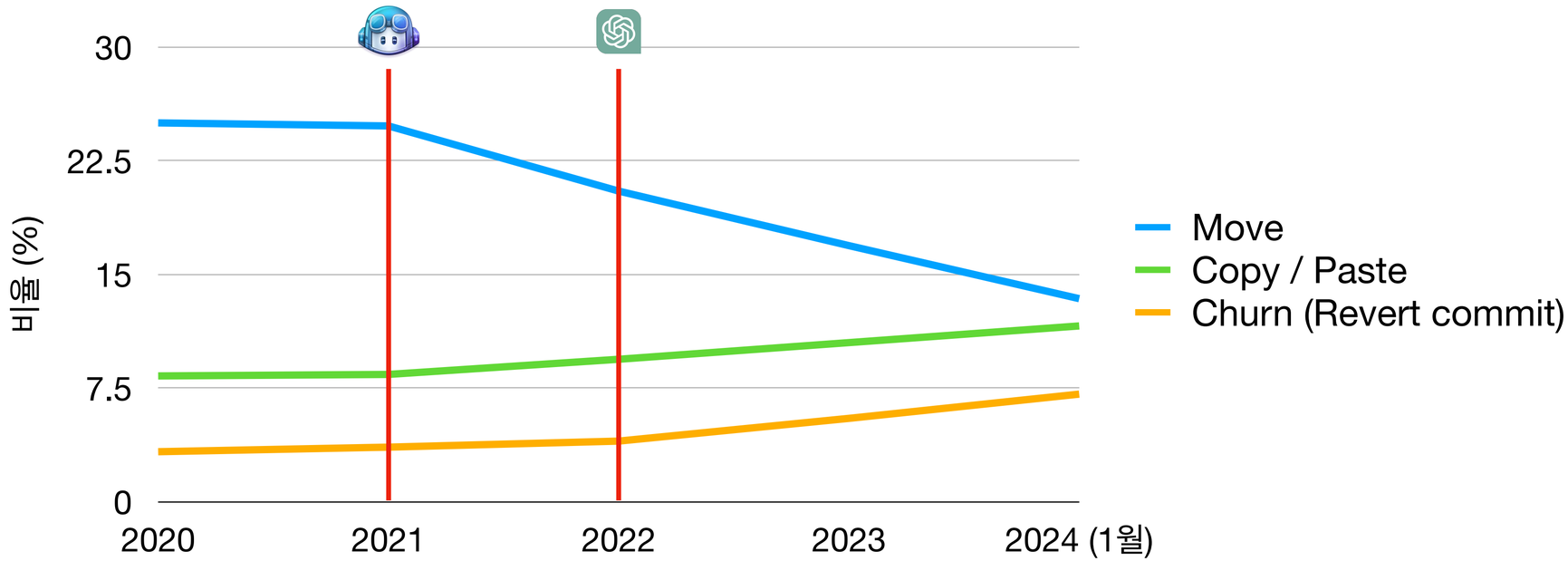


- **Refactoring** poorly

*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

Importance of interpreting LLM

- Code quality has been declining since the language model came out*.



- **Refactoring** poorly
- Not writing the correct code the first time

*Coding on Copilot 2023 Data Shows Downward Pressure on Code Quality, William Harding, Matthew Kloster, 20240116

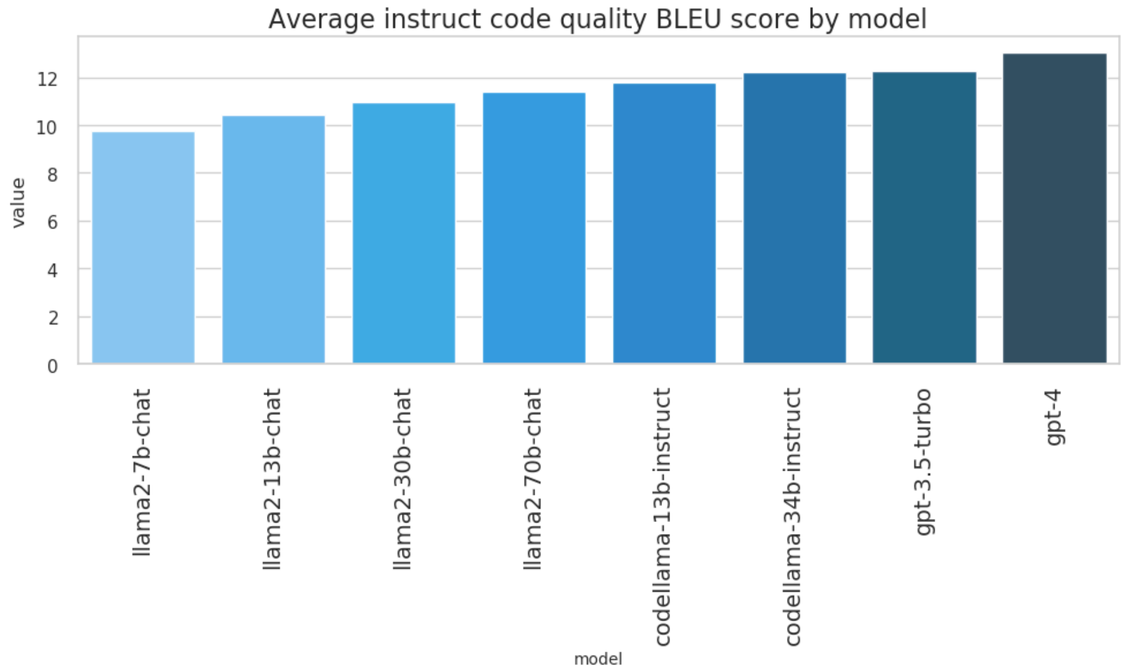
Importance of interpreting LLM

- Language models produce insecure code as **performance increases**

*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

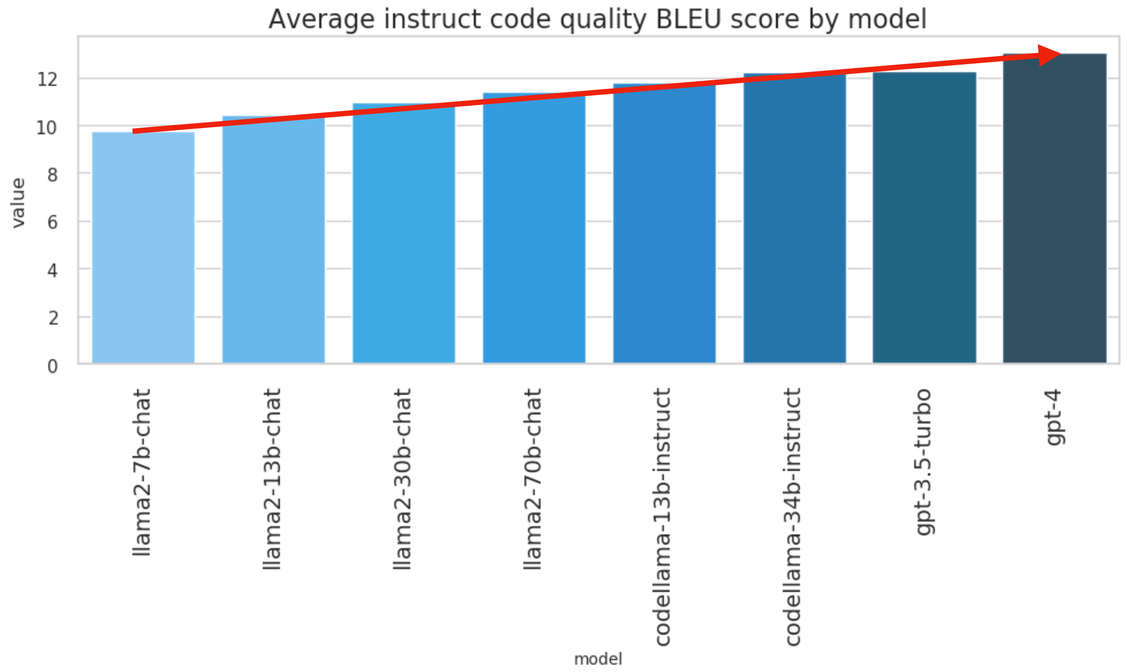
- Language models produce insecure code as **performance increases**



*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

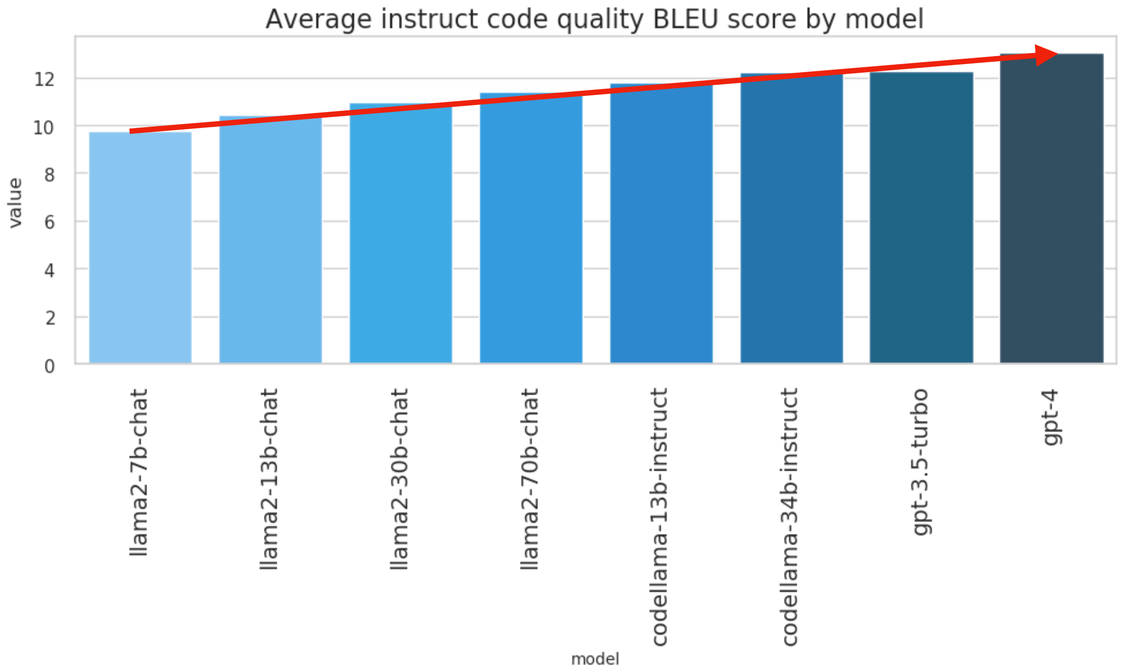
- Language models produce insecure code as **performance increases**

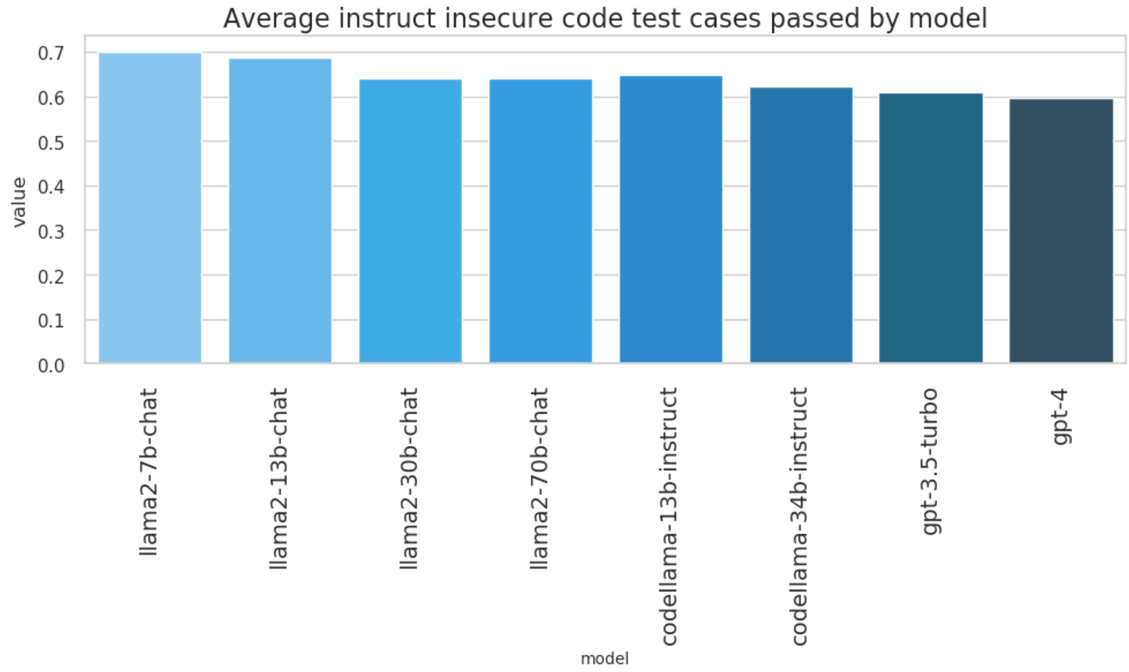


*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

- Language models produce insecure code as **performance increases**

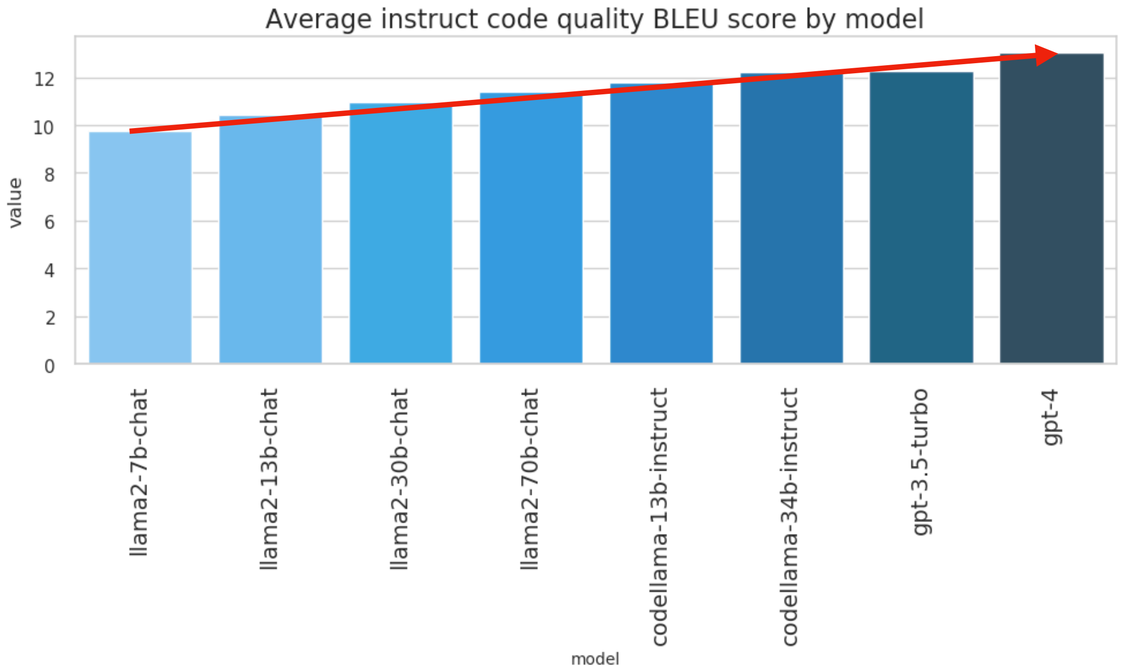


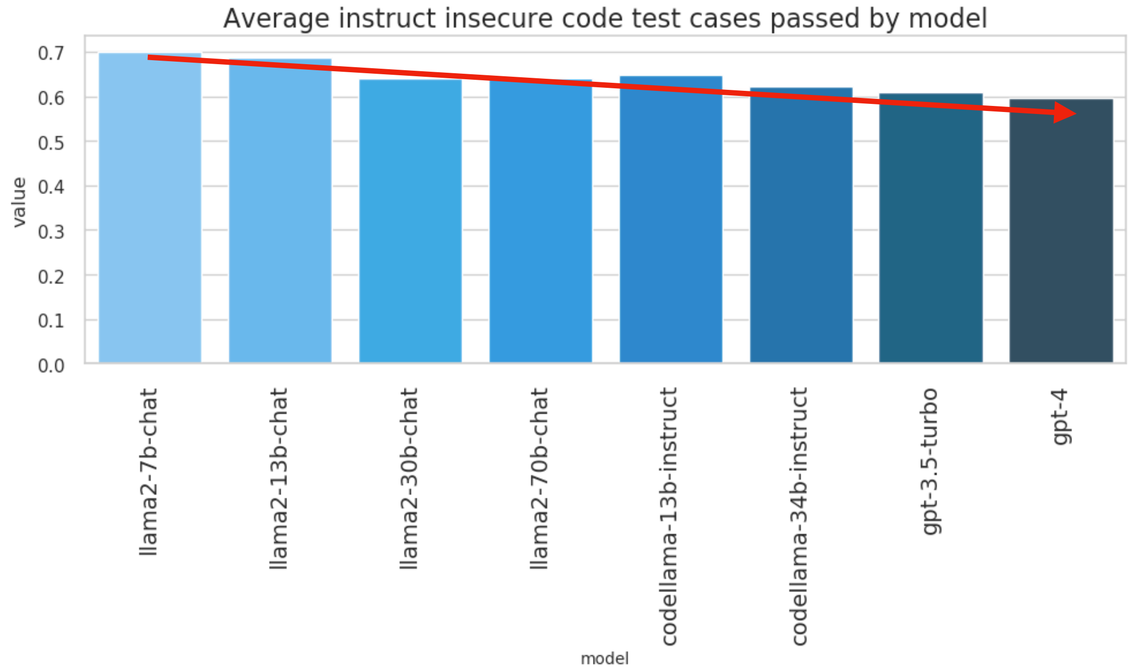


*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

- Language models produce insecure code as **performance increases**

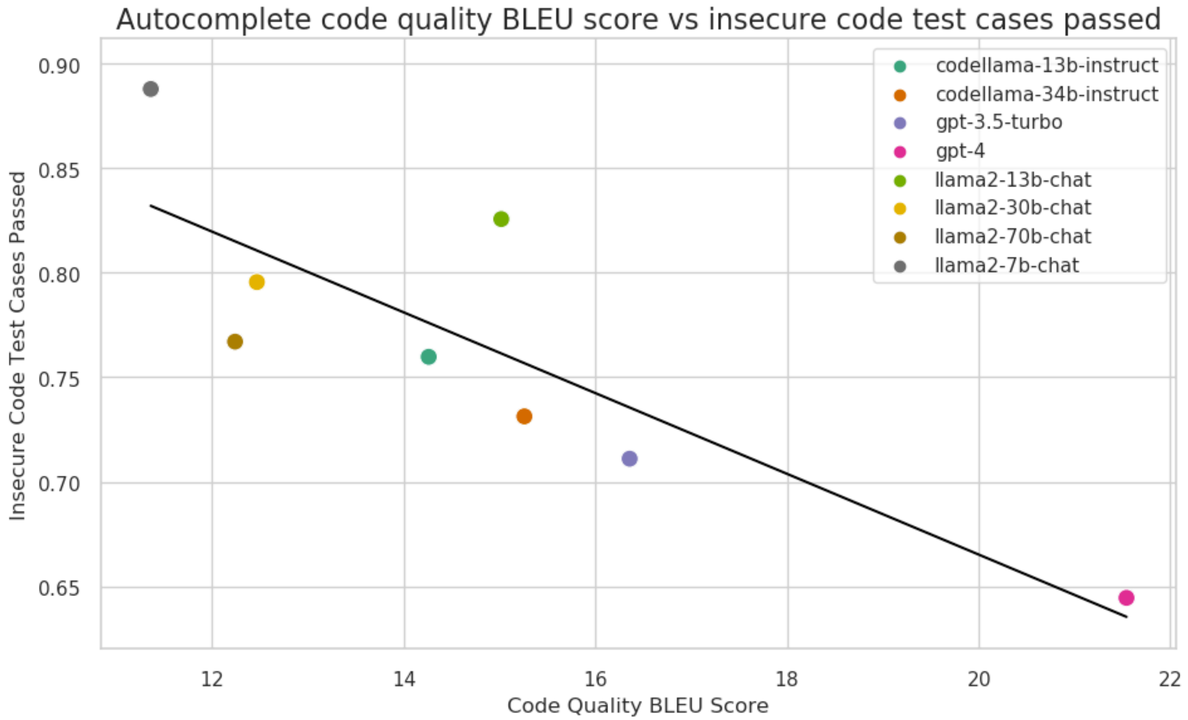


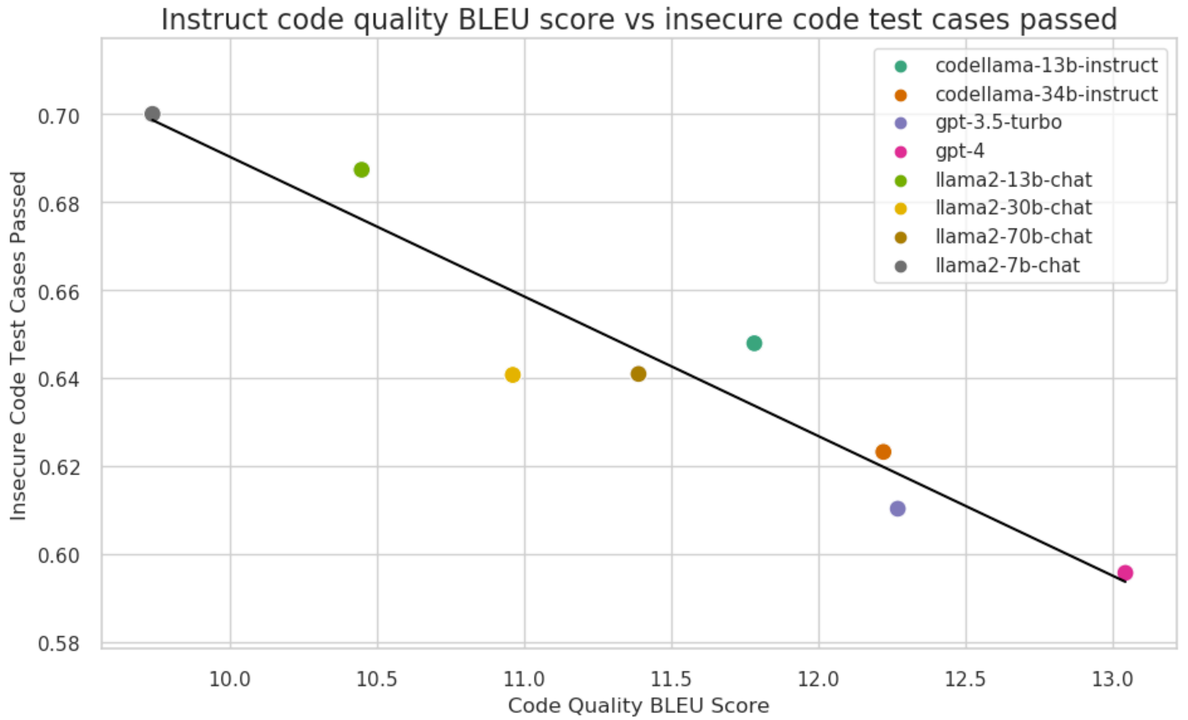


*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

- Language models produce insecure code as **performance increases**

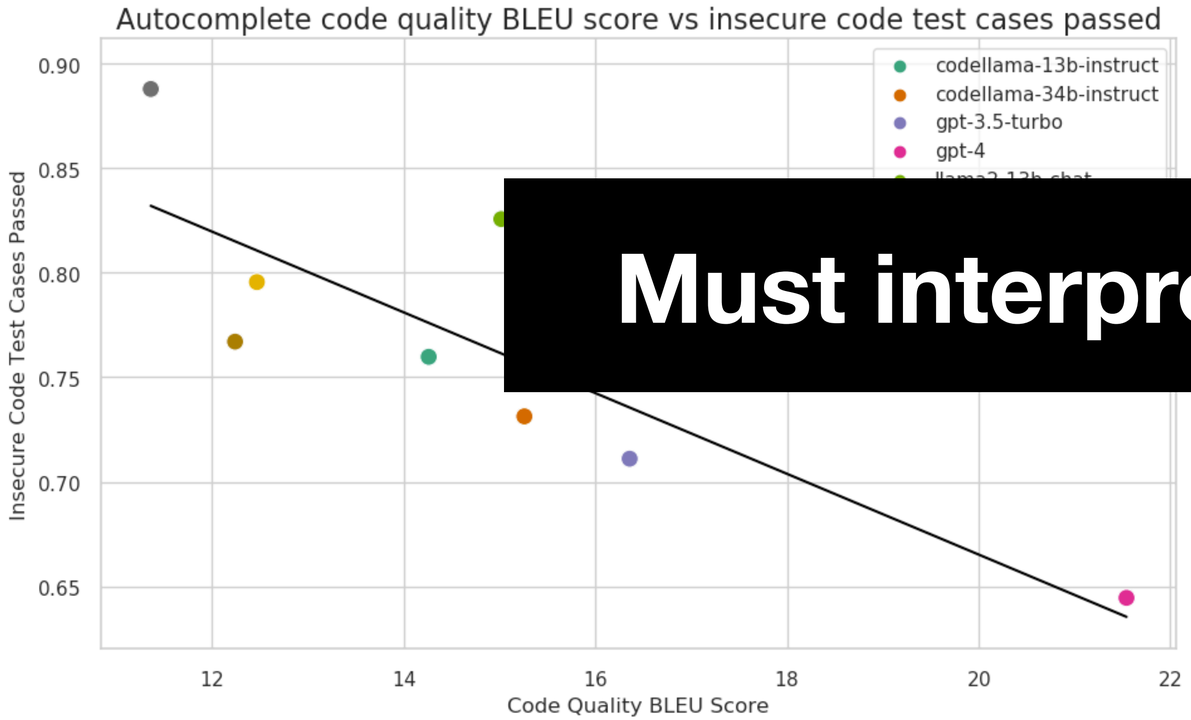


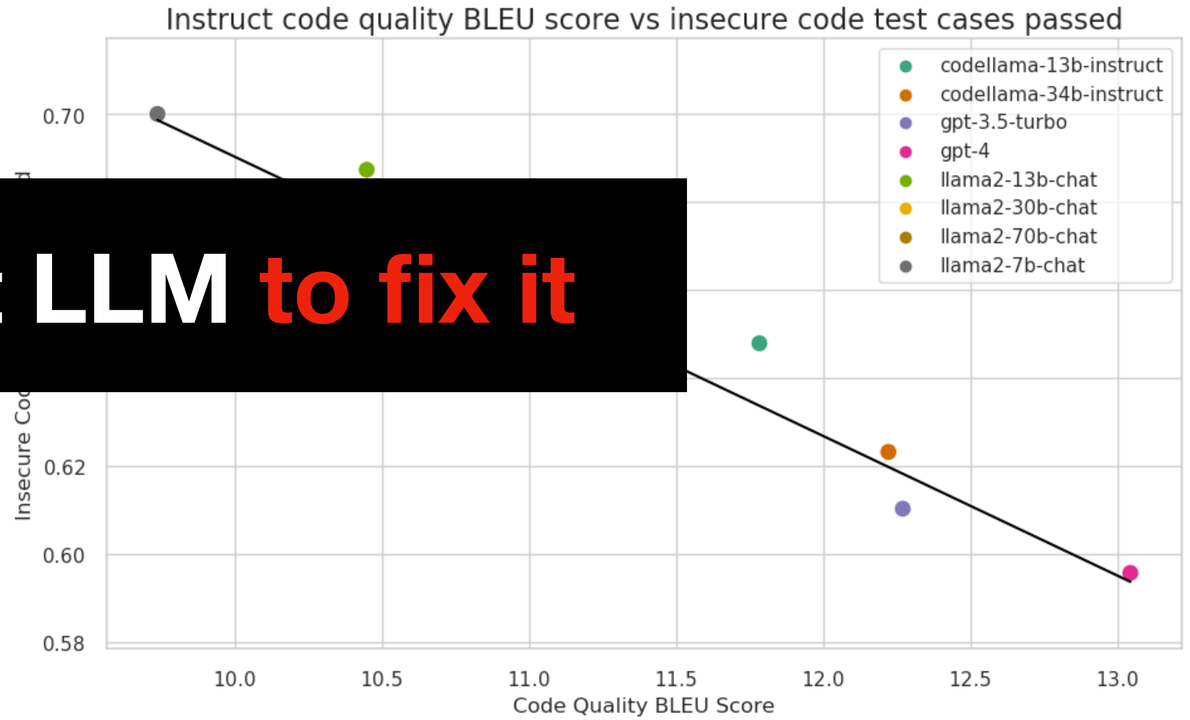


*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

- Language models produce insecure code as **performance increases**





Must interpret LLM to fix it

*Manish Bhatt et al. (2023). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models.

Importance of interpreting LLM

- But.. LLMs are **black box**

Large Language Model

Importance of interpreting LLM

- But.. LLMs are **black box**

Large Language Model

x Impossible to fix

Importance of interpreting LLM

- But.. LLMs are **black box**

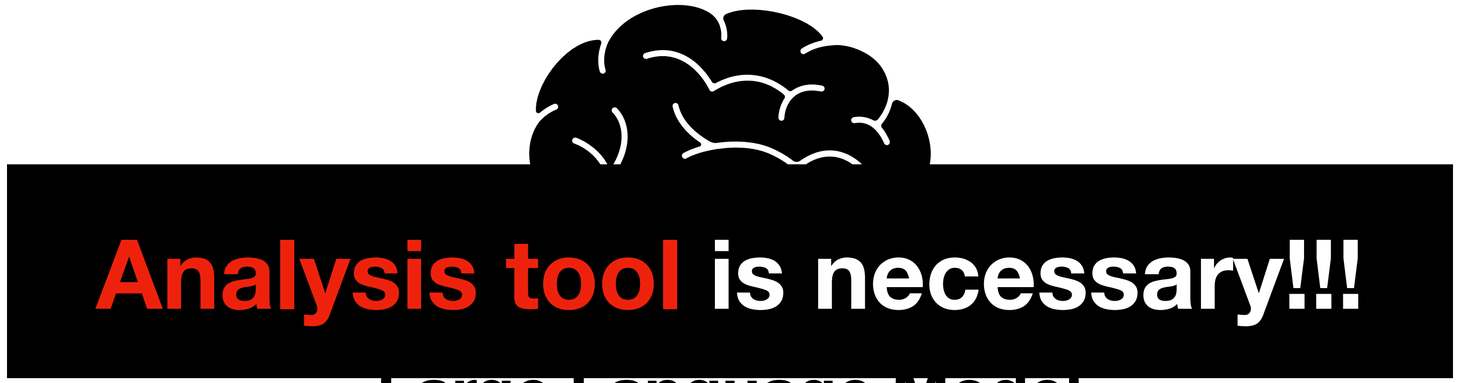
Large Language Model

x Impossible to fix

x Impossible to verify

Importance of interpreting LLM

- But.. LLMs are **black box**



x Impossible to fix

x Impossible to verify

How to interpret black box

- Binary code

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …
    - IDA Pro, Ghidra, GDB, ...

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …
    - IDA Pro, Ghidra, GDB, ...
- Language model’s prediction

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …
    - IDA Pro, Ghidra, GDB, ...
- Language model’s prediction
  - Utilizing **calculation patterns** (Attention)

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …
    - IDA Pro, Ghidra, GDB, ...
- Language model’s prediction
  - Utilizing **calculation patterns** (Attention)
  - **Code input** (Structured)

How to interpret black box

- Binary code
  - Utilizing **instruction patterns**
  - Static / Dynamic Analyzer, Decompiler, …
    - IDA Pro, Ghidra, GDB, ...
- Language model’s prediction
  - Utilizing **calculation patterns** (Attention)
  - **Code input** (Structured)
  - **Atten-Scope**

Demo

Problems of previous tools

- There are lots of tools for visualizing attention

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- There are lots of tools for visualizing attention
  - Naïve heatmap using Python

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- There are lots of tools for visualizing attention
  - Naïve heatmap using Python
  - BertViz*

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- There are lots of tools for visualizing attention
  - Naïve heatmap using Python
  - BertViz*
  - CircuitsViz**

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- There are lots of tools for visualizing attention
  - Naïve heatmap using Python
  - BertViz*
  - CircuitsViz**
  - AttentionViz***

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input
  - Bad user experience

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input
  - Bad user experience
- Not compatible with **code data**

*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input
  - Bad user experience
- Not compatible with **code data**
  - Specialized in vision models

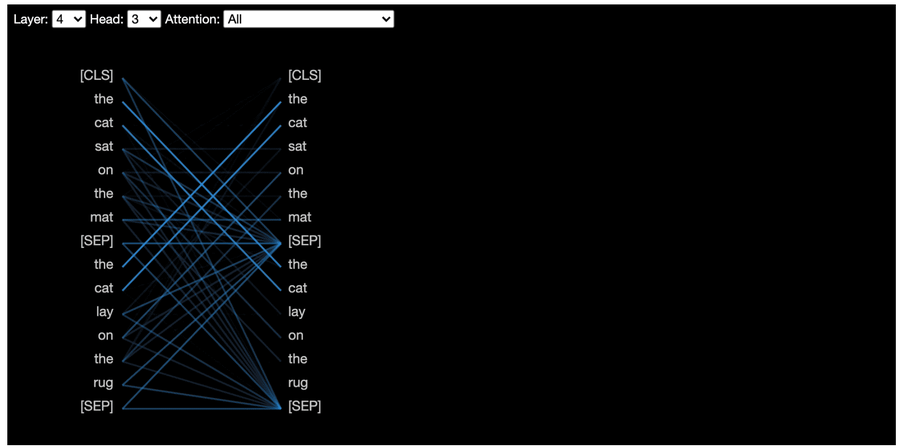
*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input
  - Bad user experience
- Not compatible with **code data**
  - Specialized in vision models



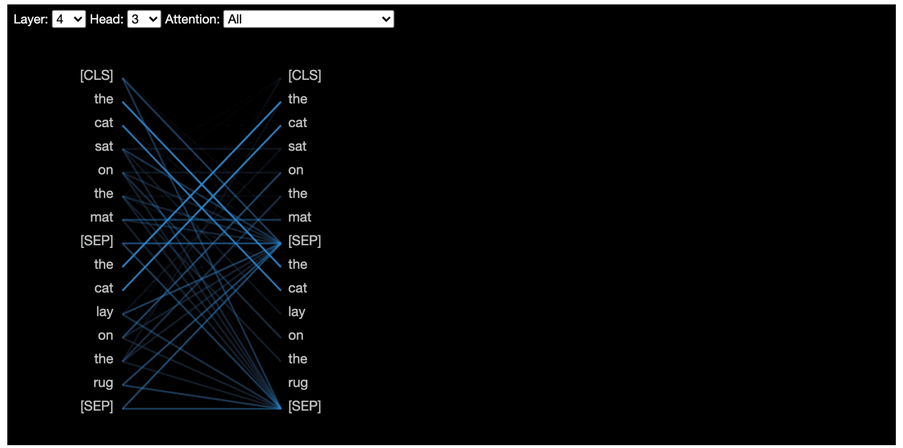
*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

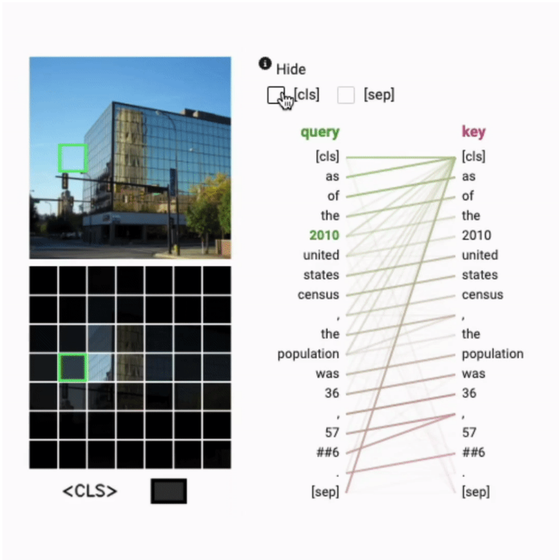
**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

Problems of previous tools

- **Not insightful**
  - **Serve noisy information**
- Too **slow**
  - We cannot analyze large-scale input
  - Bad user experience
- Not compatible with **code data**
  - Specialized in vision models





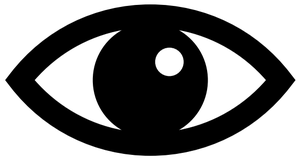
*Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics:

**https://github.com/alan-cooney/CircuitsVis

***Yeh, C., Chen, Y., Wu, A., Chen, C., Viegas, F., & Wattenberg, M. (2023). Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics.

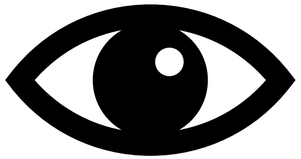
Atten-Scope

Atten-Scope

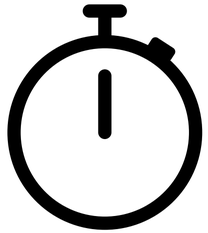


Insightful

Atten-Scope

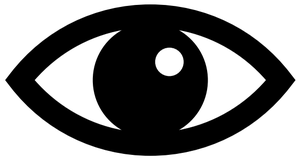


Insightful

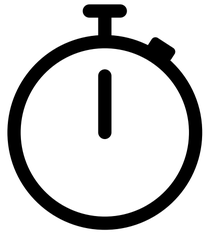


Fast

Atten-Scope



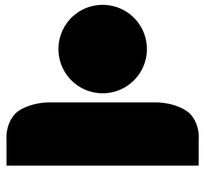
Insightful



Fast

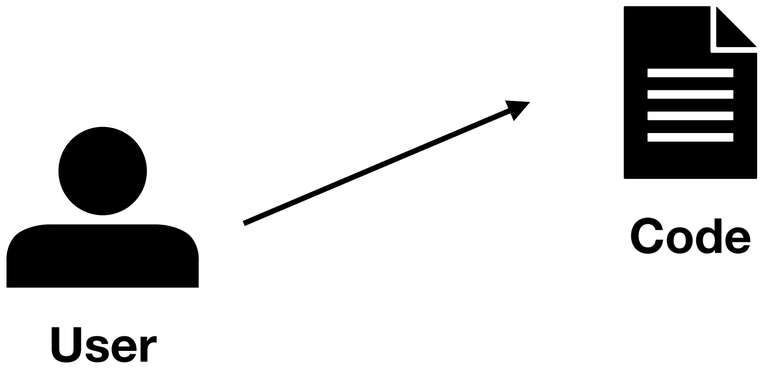
Code-friendly

Overview



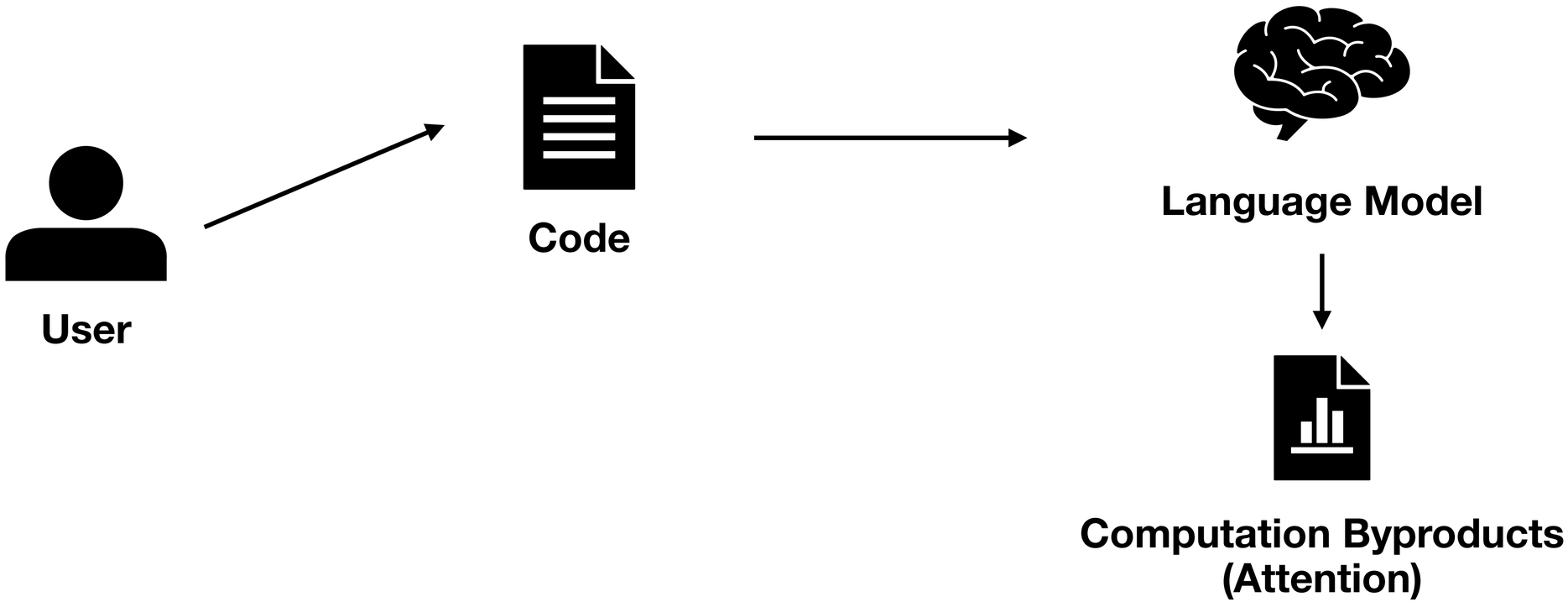
User

Overview

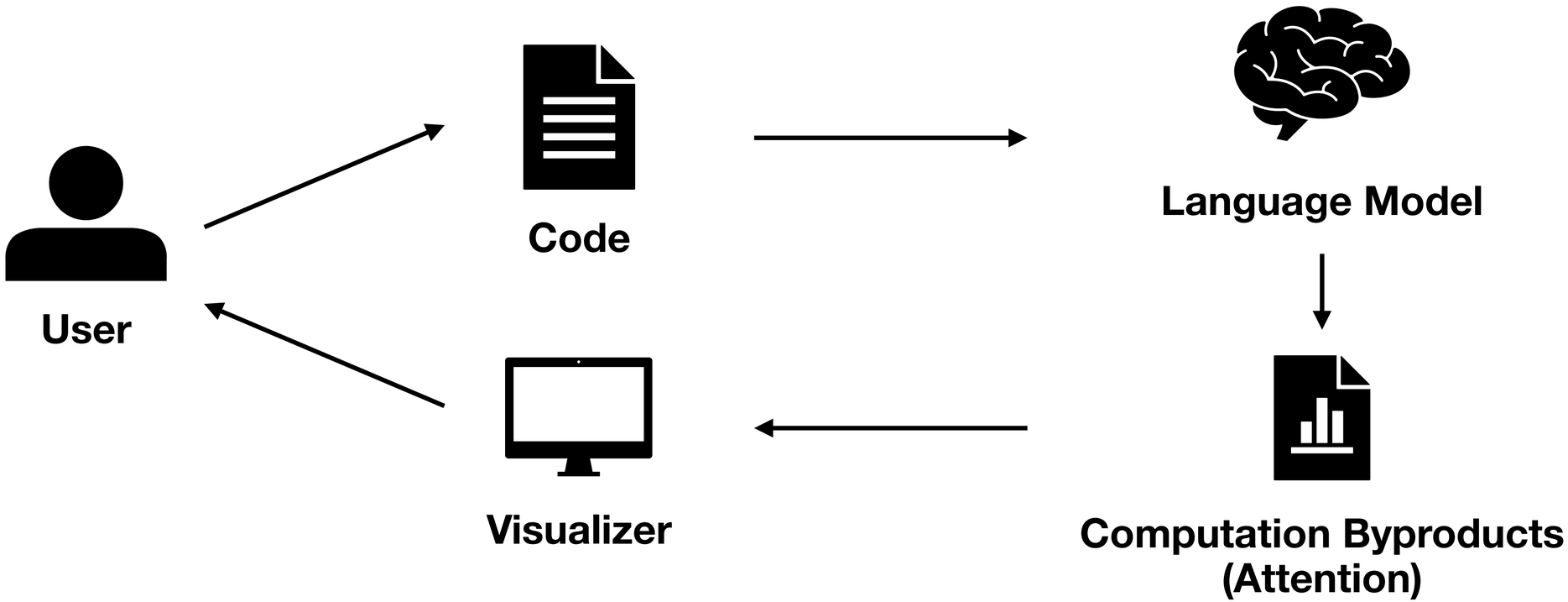


Overview

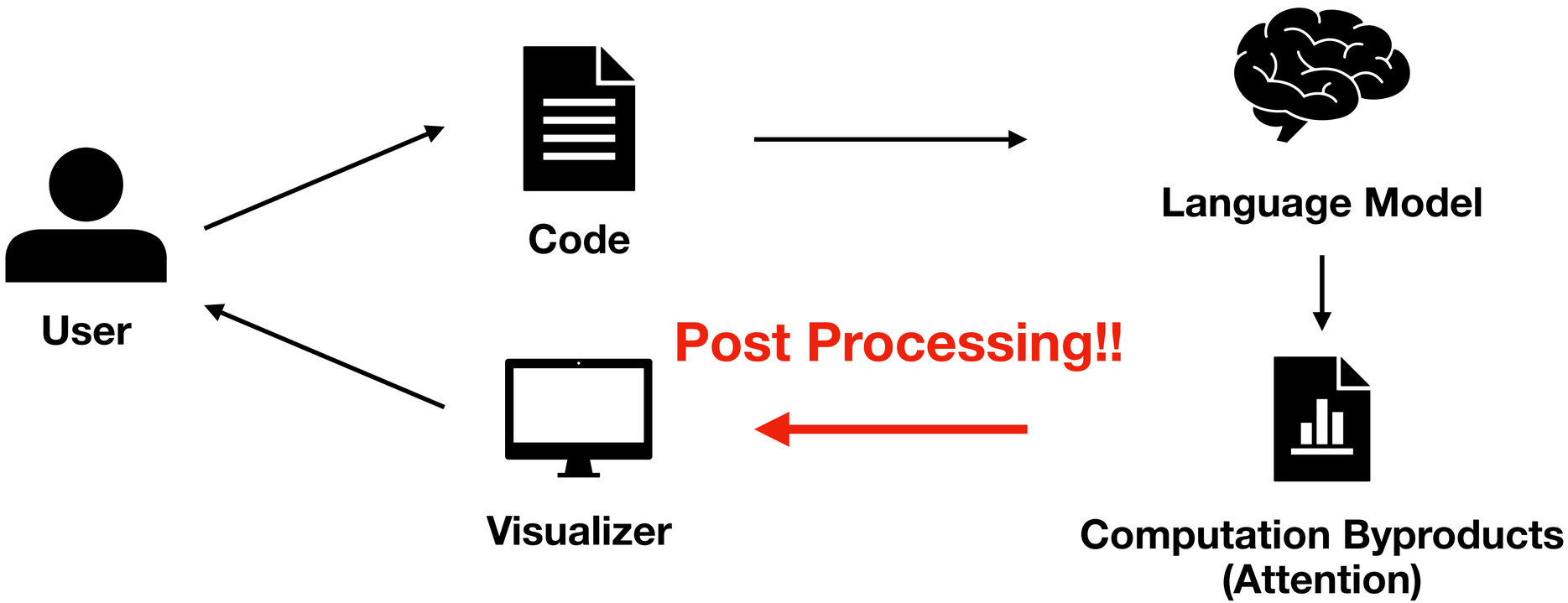
Overview


User
Code
Language Model
Computation Byproducts
(Attention)

Overview


User
Code
Language Model
Visualizer
Computation Byproducts
(Attention)

Overview


User
Code
Language Model
Post Processing!!
Visualizer
Computation Byproducts
(Attention)

Technical Details: How does attention work?

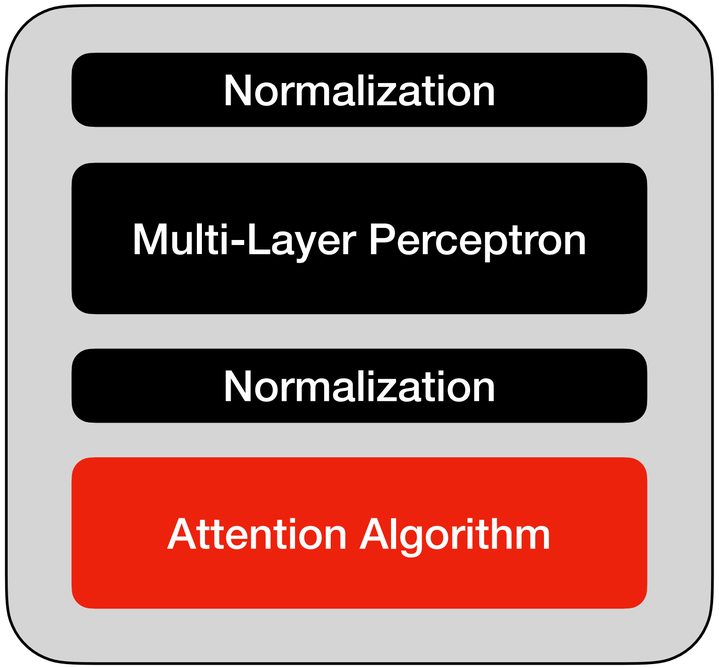
- Attention weight is an **intermediate value** for the attention algorithm!

Technical Details: How does attention work?

- Attention weight is an **intermediate value** for the attention algorithm!
- **Most of the LLMs use attention algorithm!**

Technical Details: How does attention work?

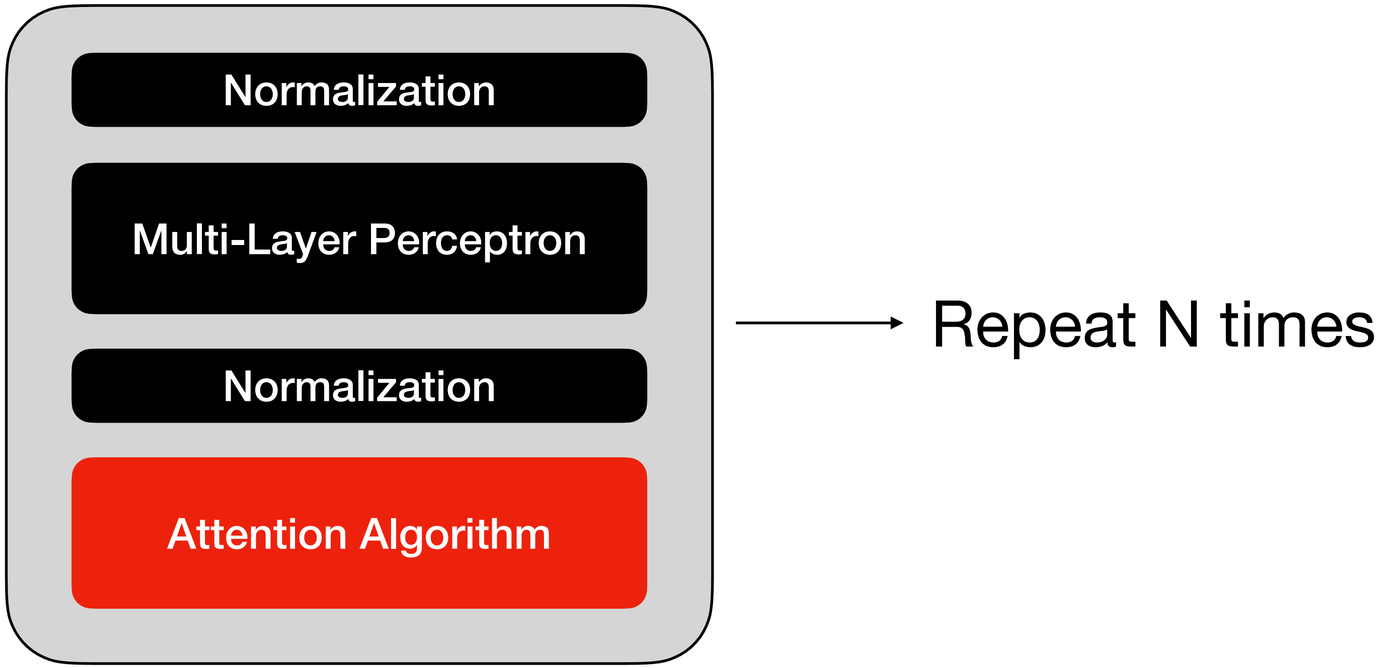
- Attention weight is an **intermediate value** for the attention algorithm!
- **Most of the LLMs use attention algorithm!**



Transformer block

Technical Details: How does attention work?

- Attention weight is an **intermediate value** for the attention algorithm!
- **Most of the LLMs use attention algorithm!**

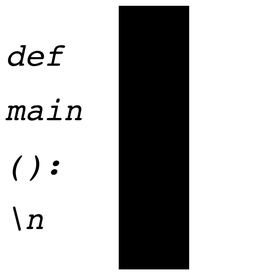


Transformer block

Technical Details: How does attention work?

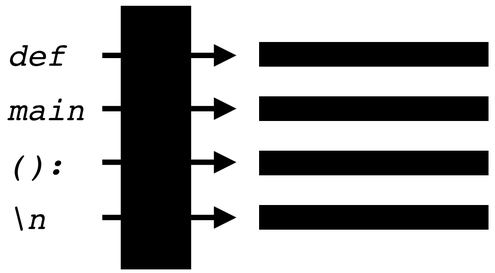
```
def
main
():
\n
```

Technical Details: How does attention work?

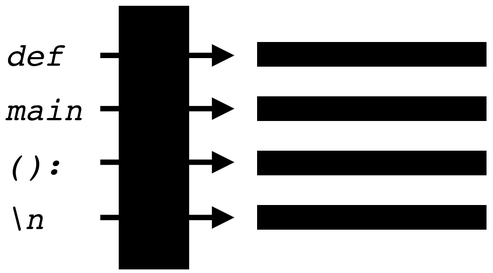


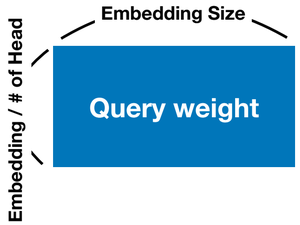
Token Encoder

Technical Details: How does attention work?

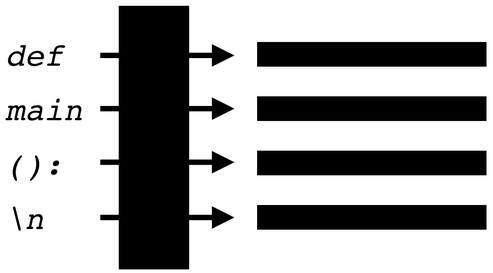


Technical Details: How does attention work?

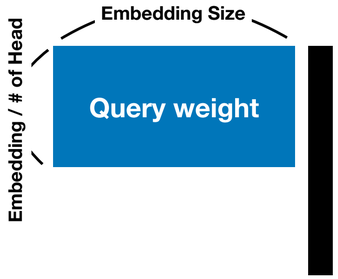




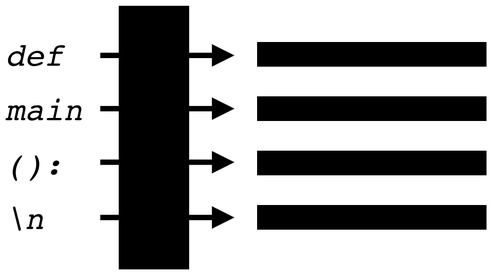
Technical Details: How does attention work?



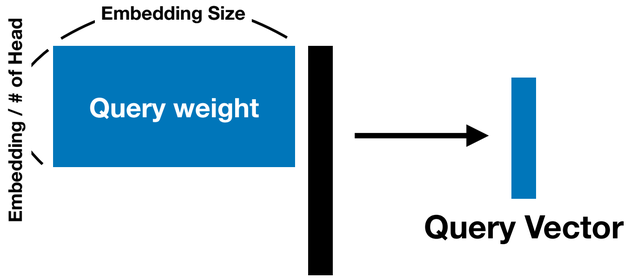
Token Encoder



Technical Details: How does attention work?

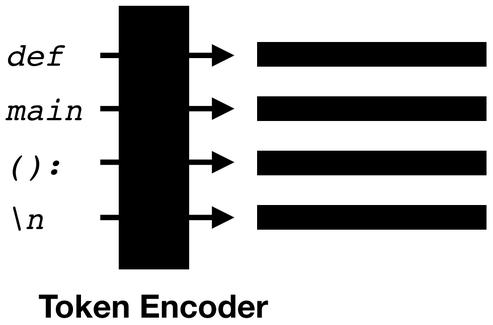


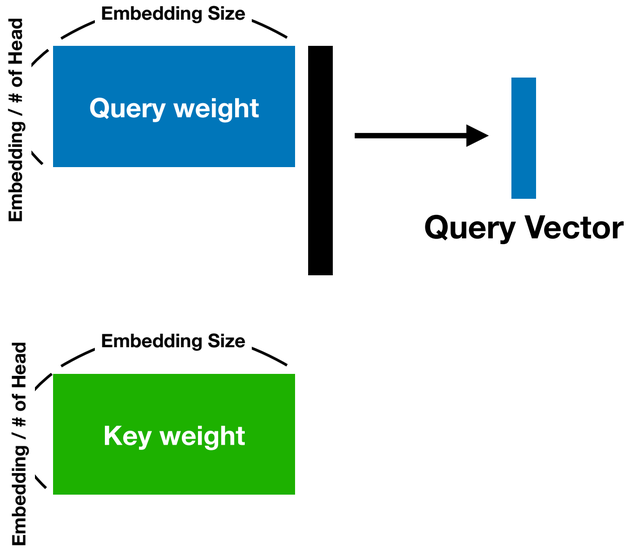
Token Encoder



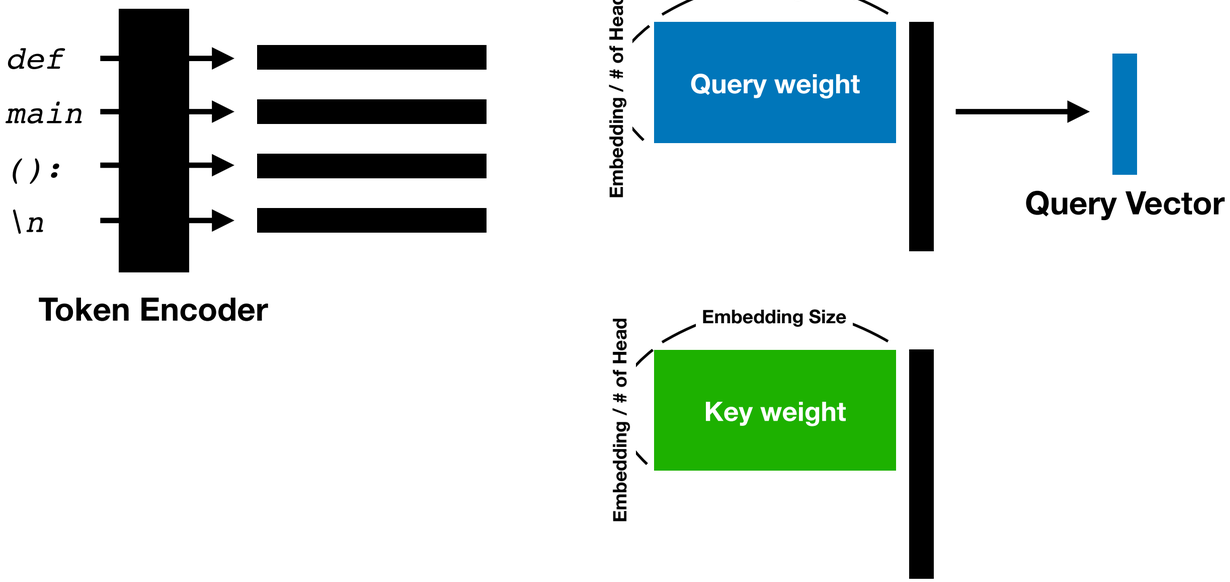
Query Vector

Technical Details: How does attention work?

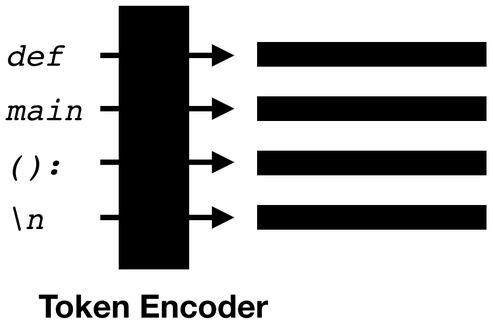


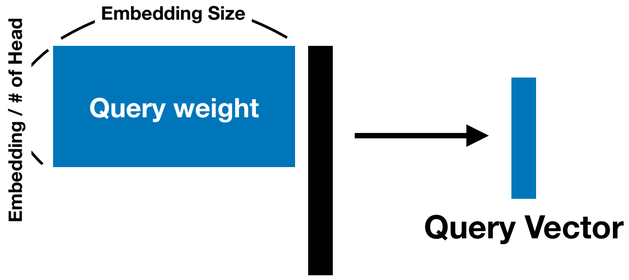


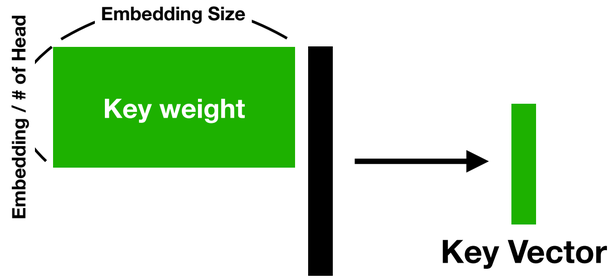
Technical Details: How does attention work?



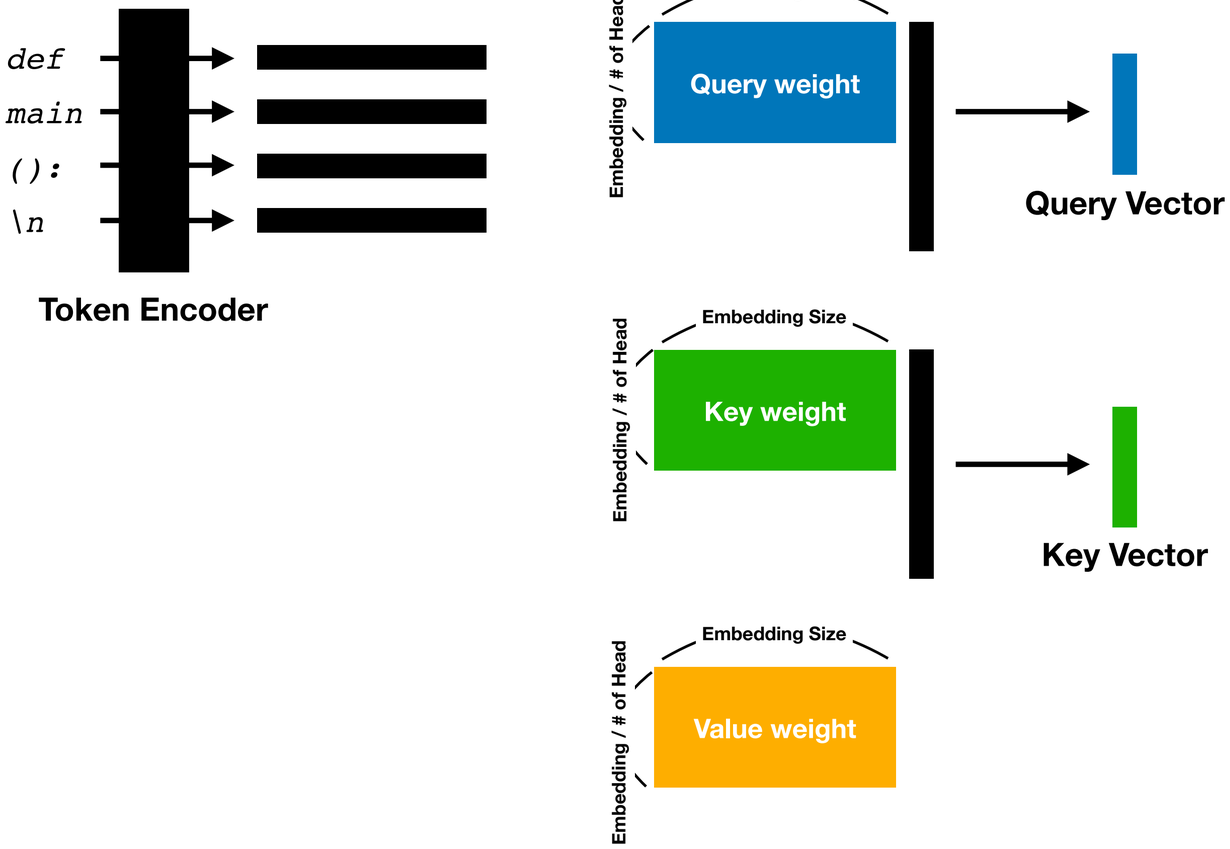
Technical Details: How does attention work?



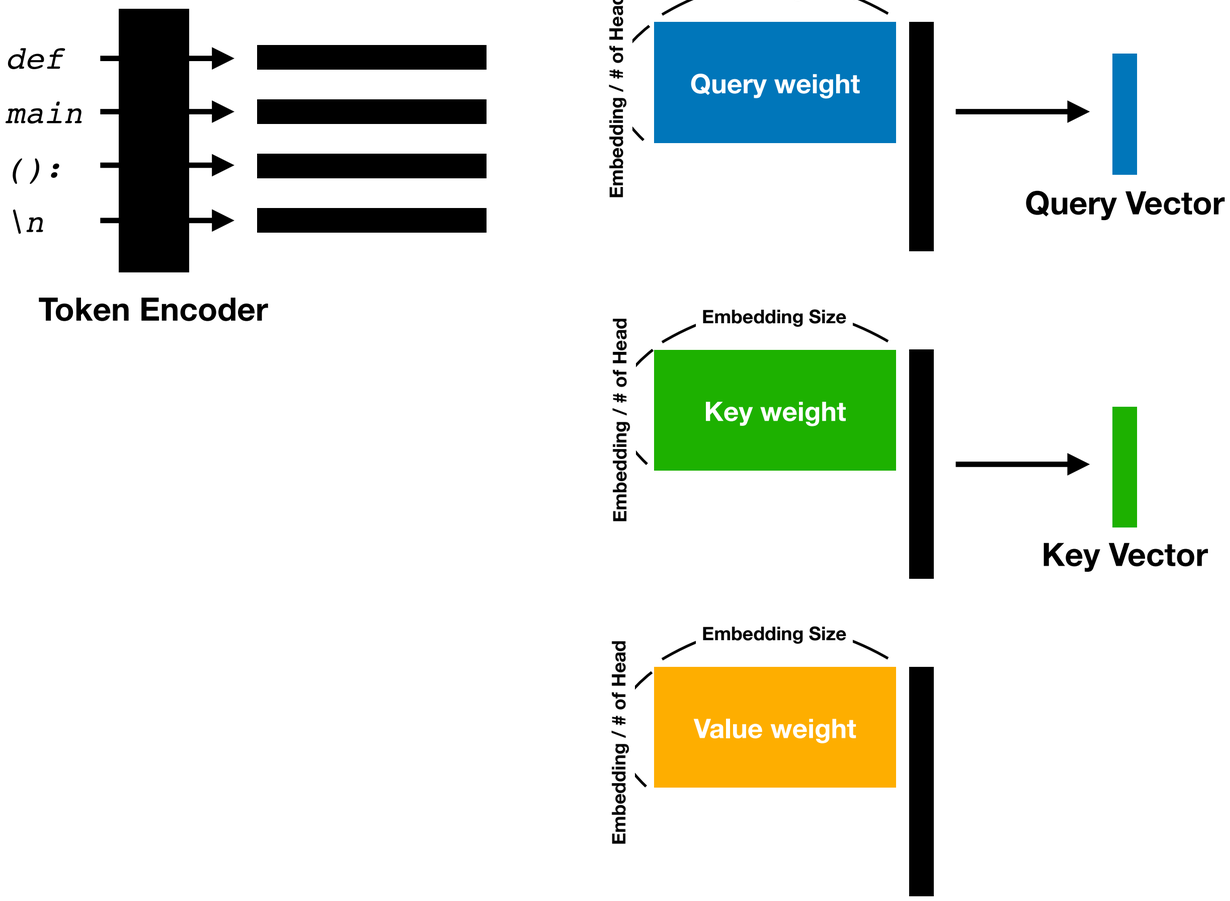




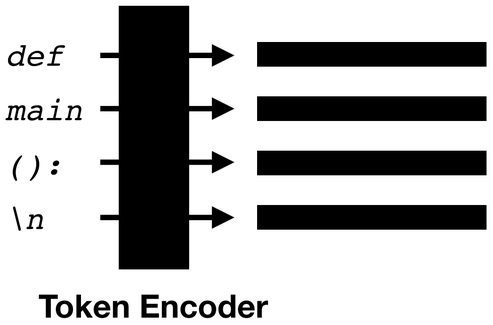
Technical Details: How does attention work?

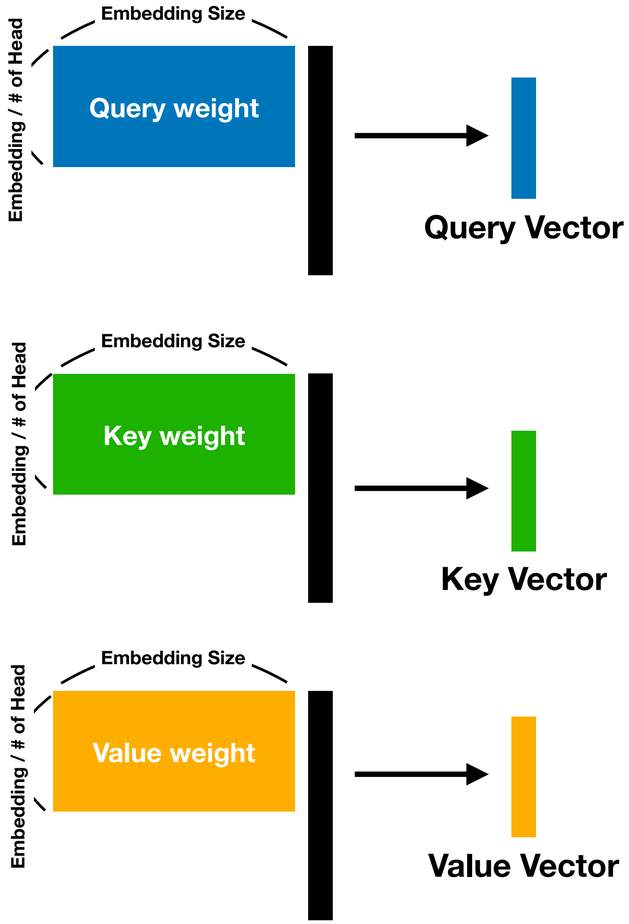


Technical Details: How does attention work?

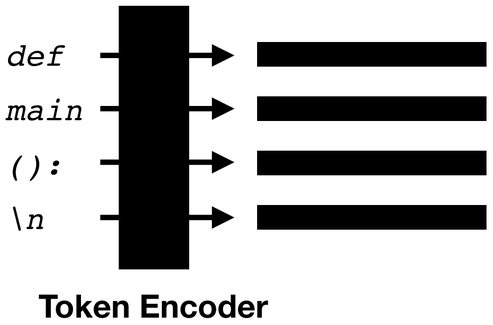


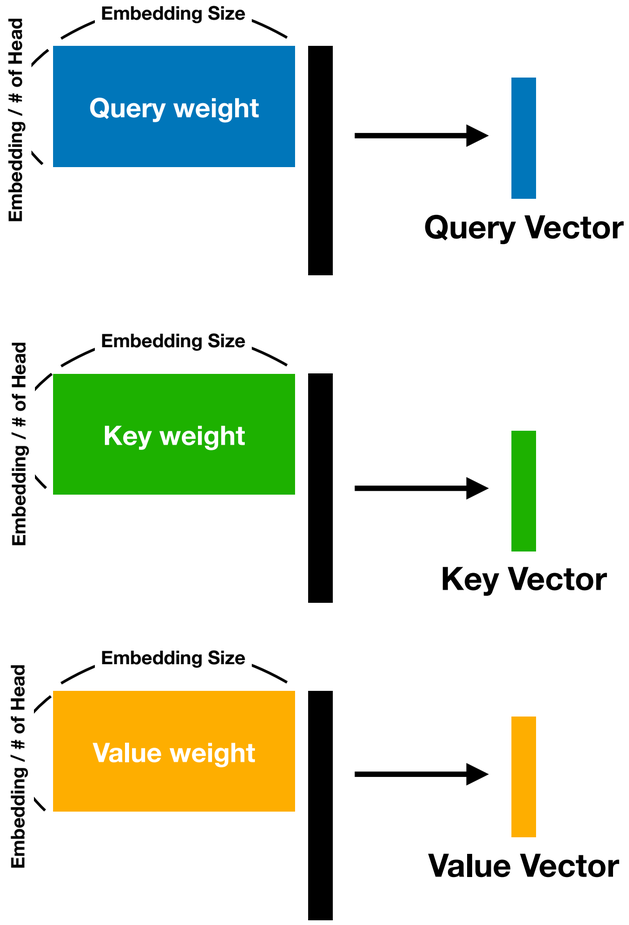
Technical Details: How does attention work?

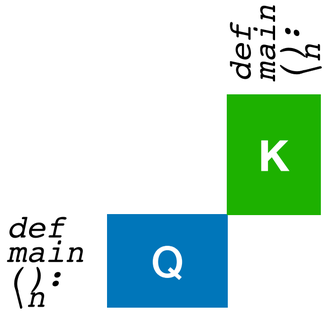




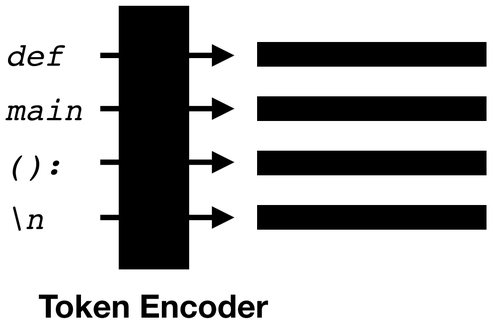
Technical Details: How does attention work?

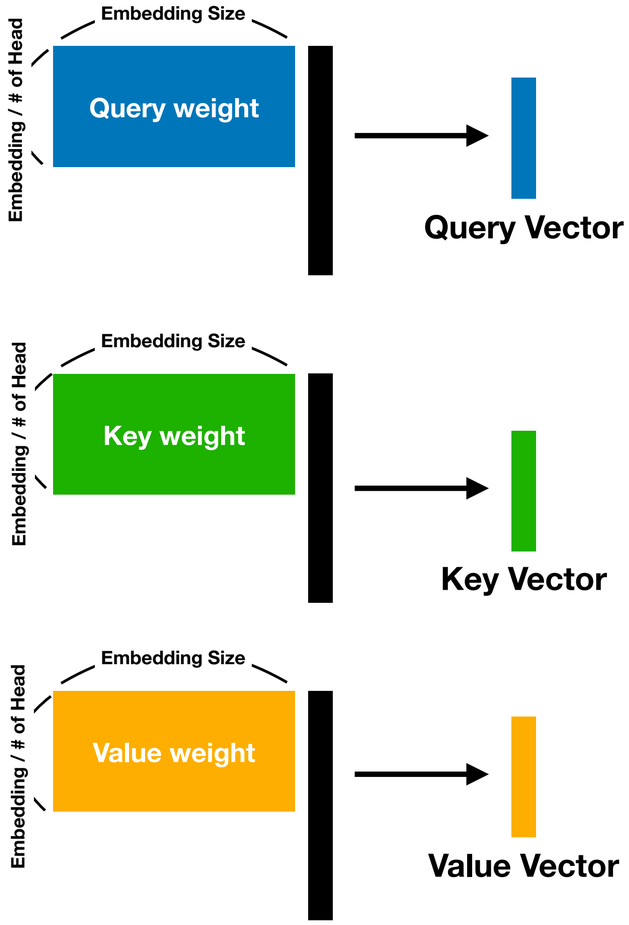


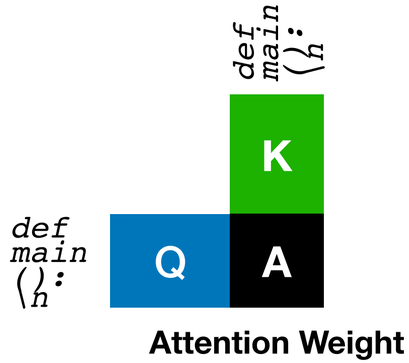




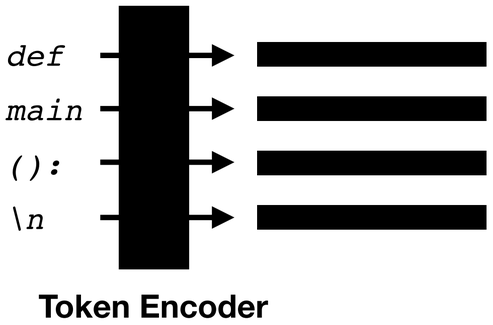
Technical Details: How does attention work?

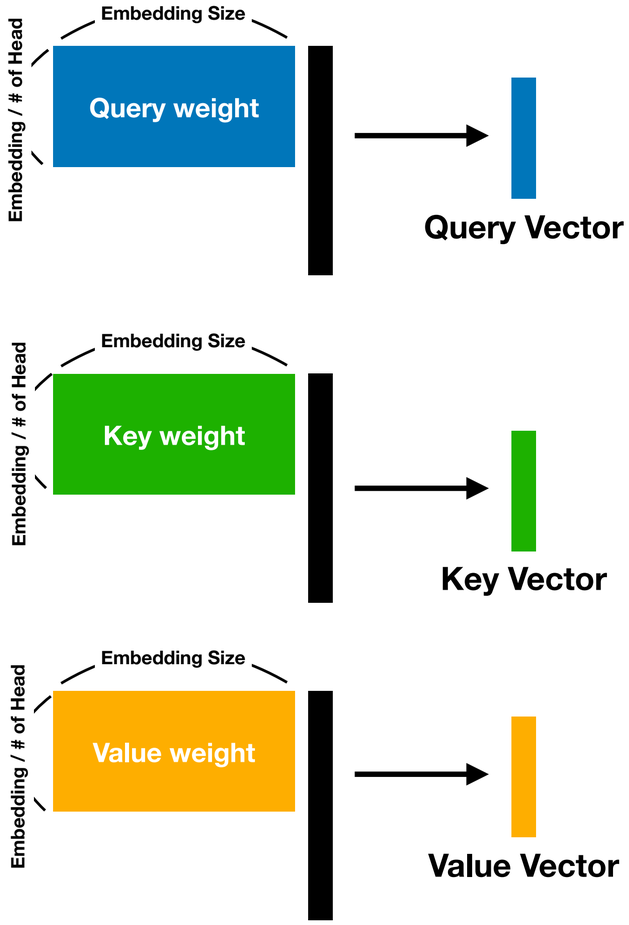


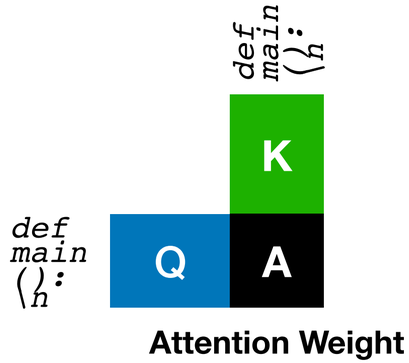




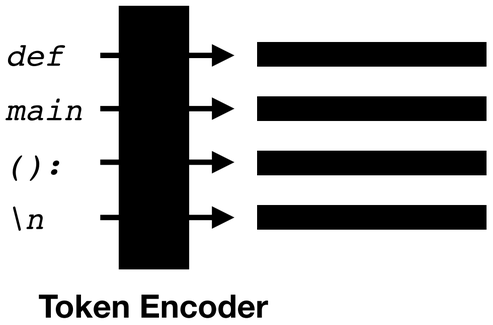
Technical Details: How does attention work?

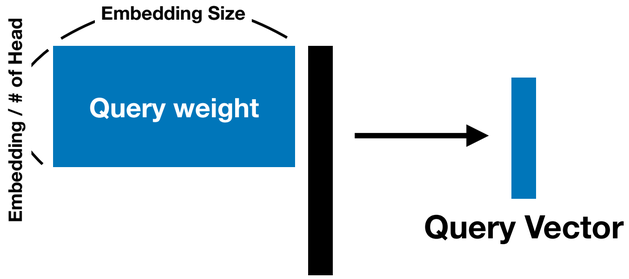


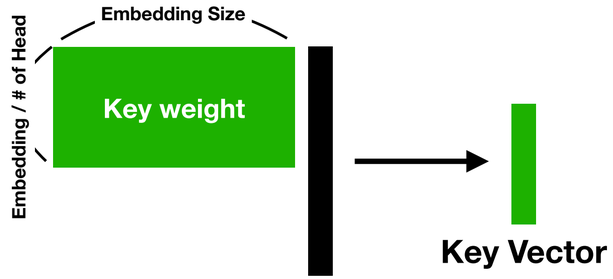


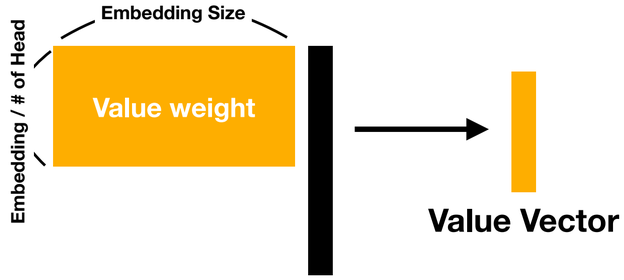


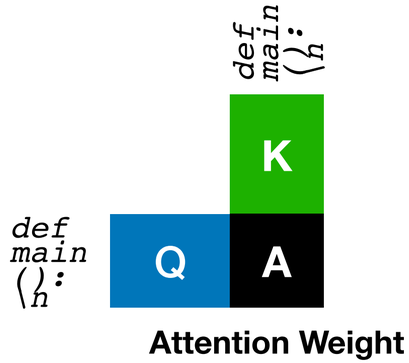
Technical Details: How does attention work?



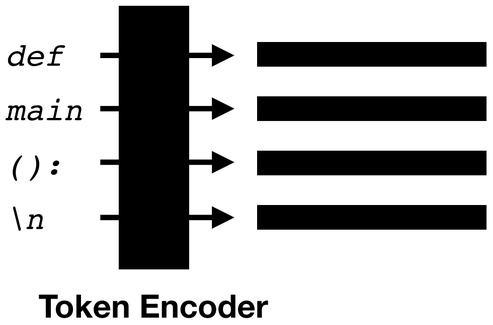


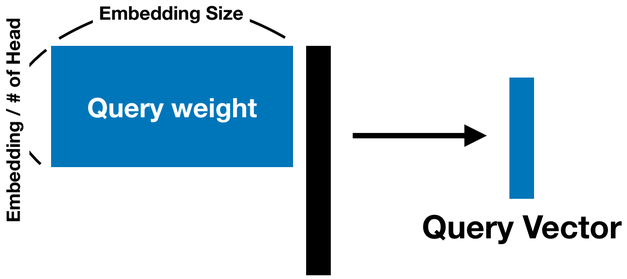


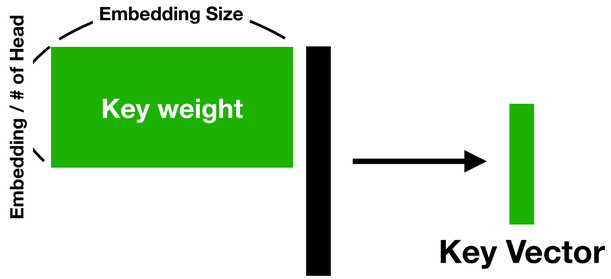


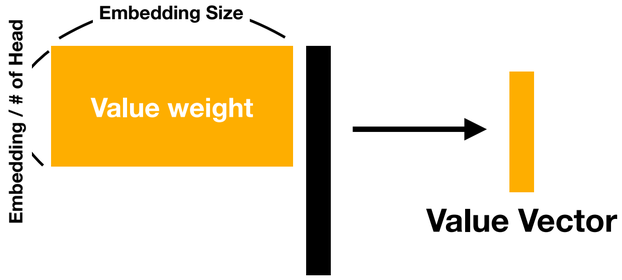


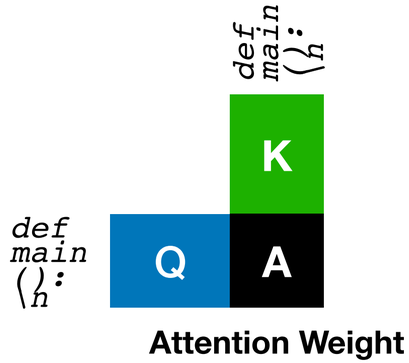
Technical Details: How does attention work?



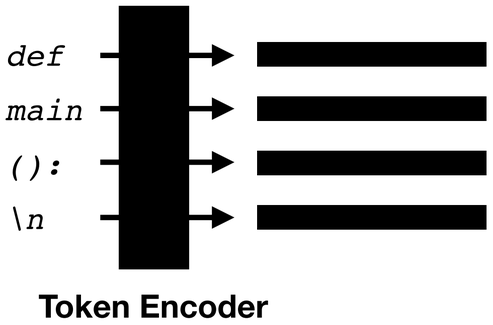


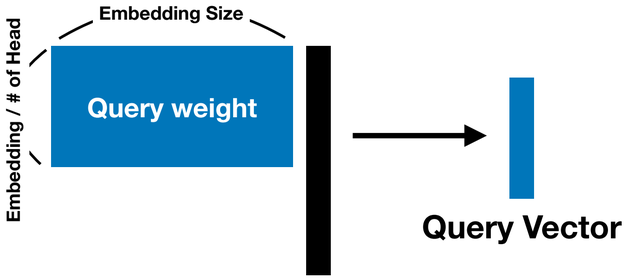


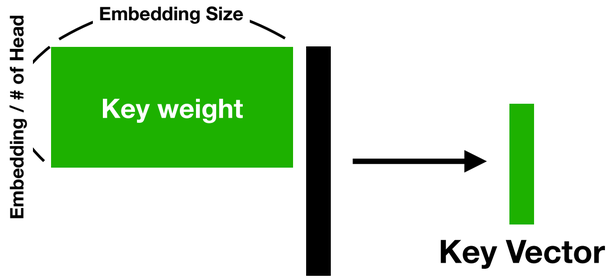


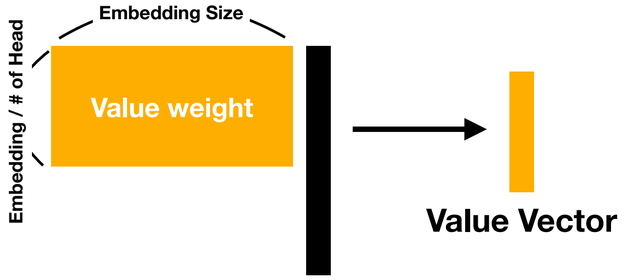


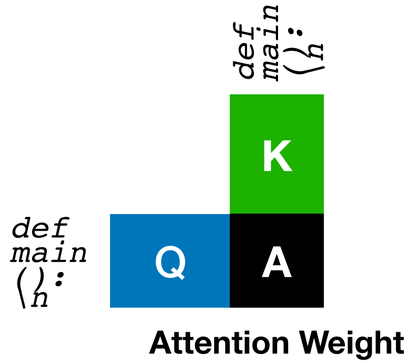
Technical Details: How does attention work?



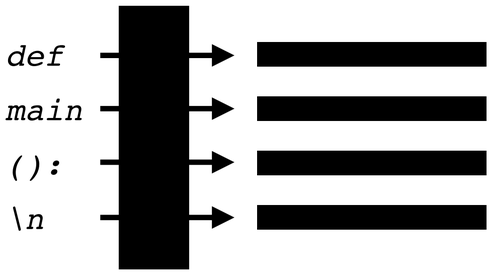




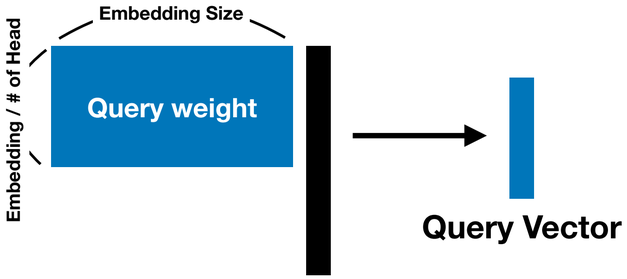




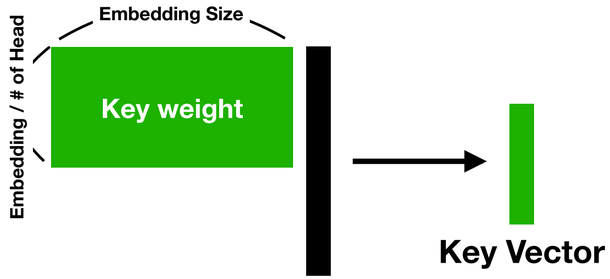
Technical Details: How does attention work?



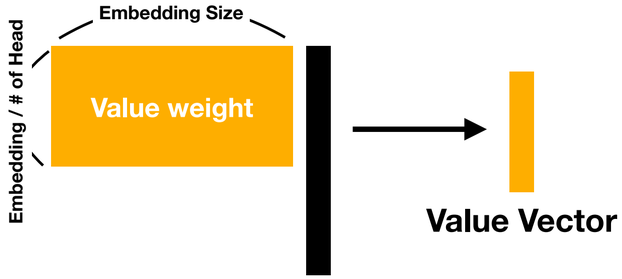
Token Encoder



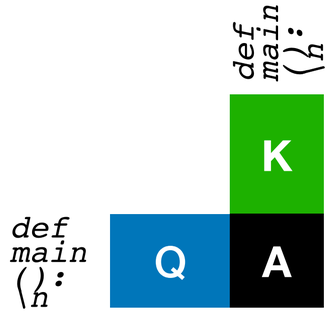
Query Vector



Key Vector



Value Vector



Attention Weight

Multi-head Attention

Technical Details: How does attention work?

	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4

Technical Details: How does attention work?

	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4

To predict the next token of `(n):`,
we need to see `fibo` as much as 0.5

Technical Details: How does attention work?

	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4

`(n):` · `fibo`

To predict the next token of `(n):`,
we need to see `fibo` as much as 0.5

Technical Details: How does attention work?

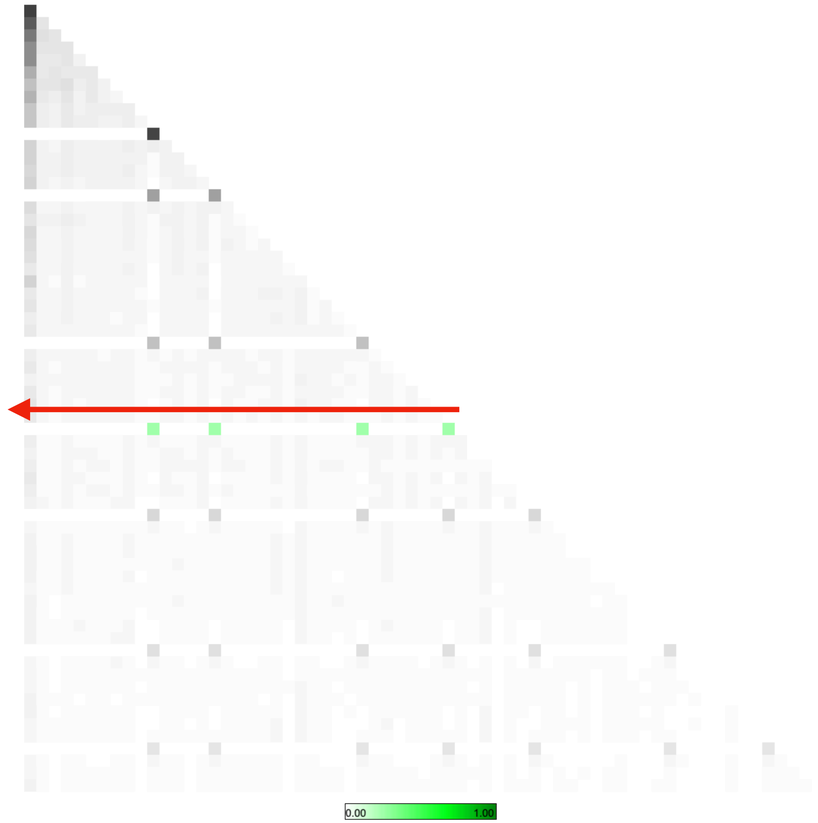
	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4

(n): · fibo ⟶ 0.5

To predict the next token of `(n):`,
we need to see `fibo` as much as 0.5

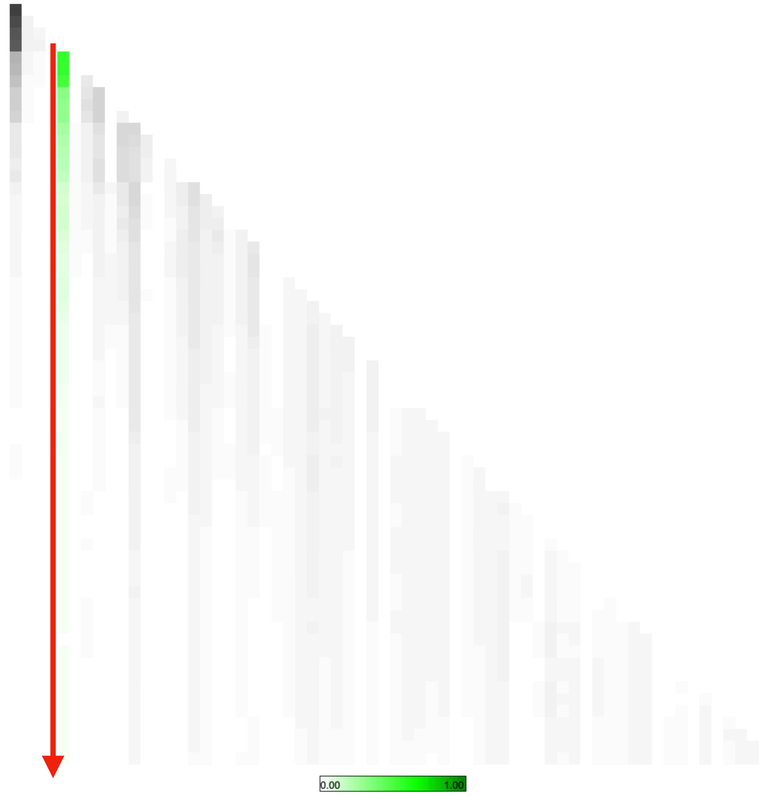
Technical Details: How does attention work?

	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4



Technical Details: How does attention work?

	def	fibo	na	cci	(n):	\n	\t	if n	==	or
def	1									
fibo	0.4	0.6								
na	0.2	0.4	0.4							
cci	0.1	0.2	0.2	0.5						
(n):	0.3	0.5	0.1	0.05	0.05					
\n	0.4	0	0	0.2	0.1	0.3				
\t	0.2	0.1	0.35	0	0.35	0	0			
if n	0	0	0.2	0.3	0.1	0.1	0.3	0		
==	0	0	0	0	0	0	0	0.6	0.4	
or	0	0	0	0	0	0	0	0.2	0.4	0.4



Technical Details: Post processing

- Raw attention is noisy

**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

- How

**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

- How
    1. Remove **null-space component***

**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

- How
  1. Remove **null-space component***
  2. Consider size of **value vector****

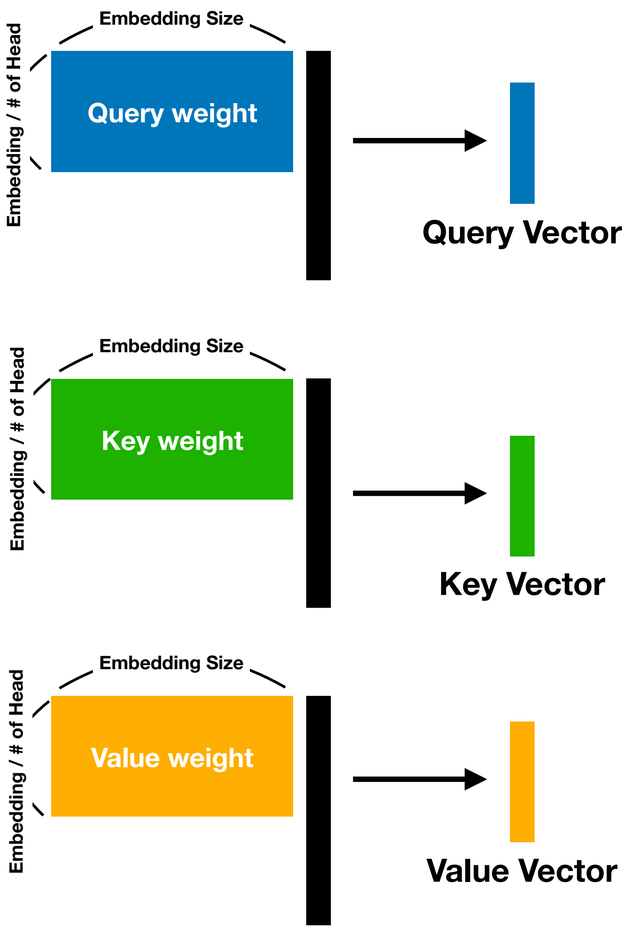
**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Technical Details: Post processing

- Raw attention is noisy
- Posting processing removes noise

- How
  1. Remove **null-space component***
  2. Consider size of **value vector****



**Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In Proceedings of the 2019 ACL Workshop BlackboxNLP

**Kobayashi, G. et al.. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In Proceedings of the 2020 EMNLP

Remove null-space component

- What is null space?

Remove null-space component

- What is null space?

Remove null-space component

- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

Remove null-space component

- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

- Null vector: A solution of some linear equation

Remove null-space component

- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

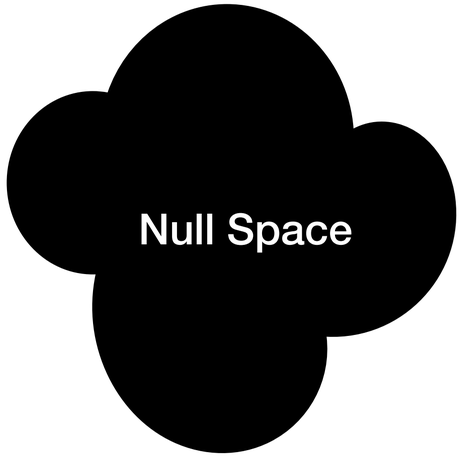
- Null vector: A solution of some linear equation
- Null space: Linear combination of solutions

Remove null-space component

- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

- Null vector: A solution of some linear equation
- Null space: Linear combination of solutions

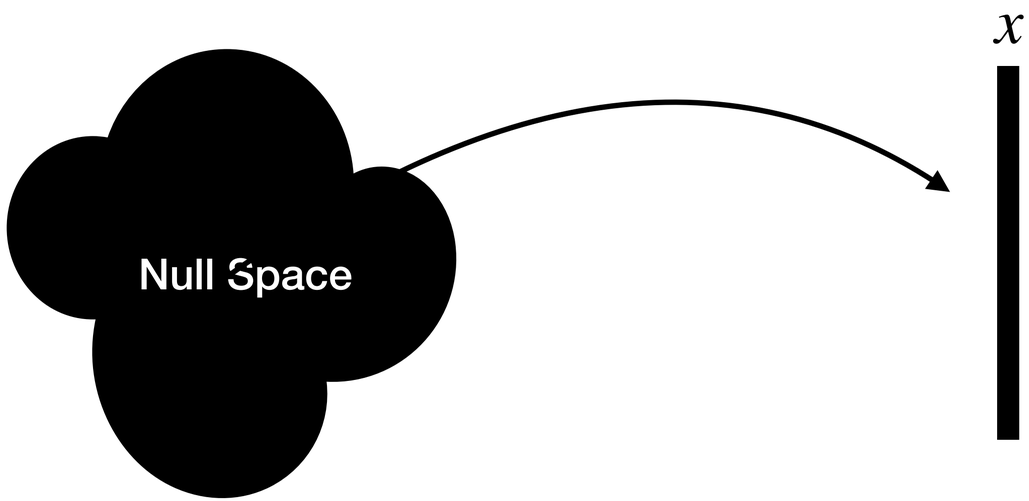


Remove null-space component

- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

- Null vector: A solution of some linear equation
- Null space: Linear combination of solutions

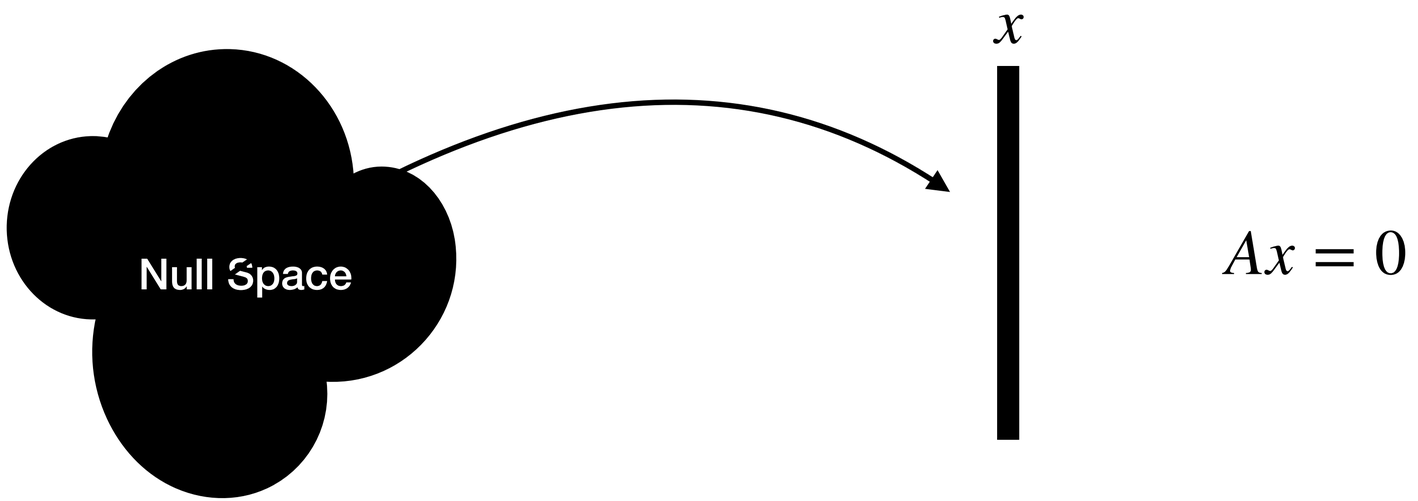


Remove null-space component

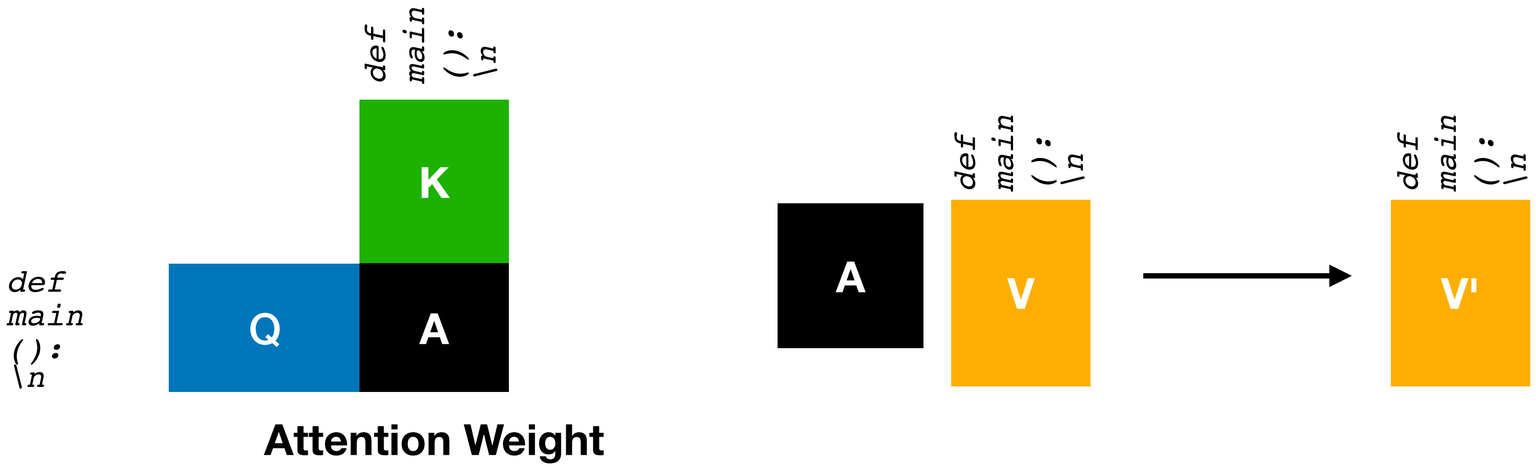
- What is null space?

$$Ax = 0 \implies x \text{ is null vector}$$

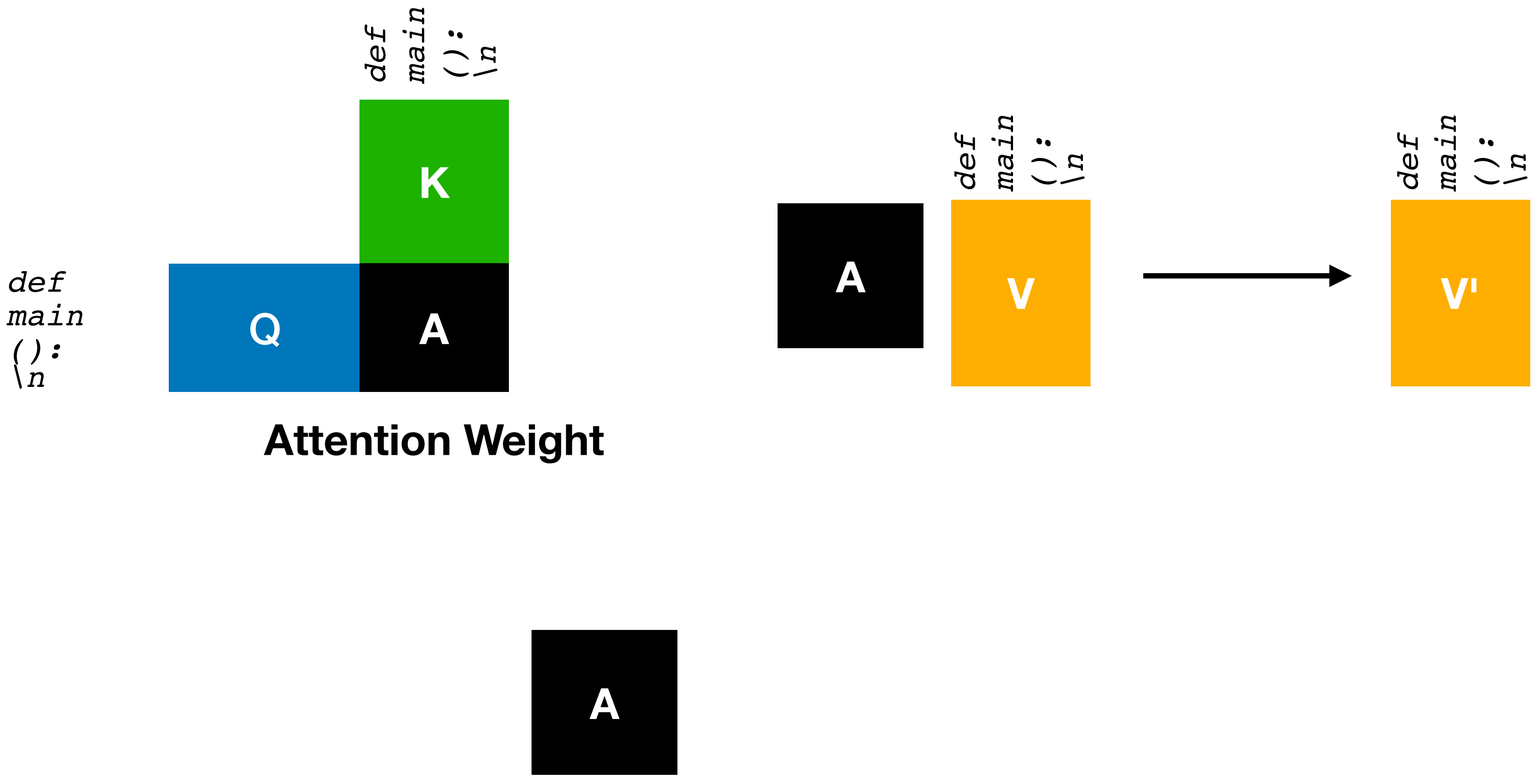
- Null vector: A solution of some linear equation
- Null space: Linear combination of solutions



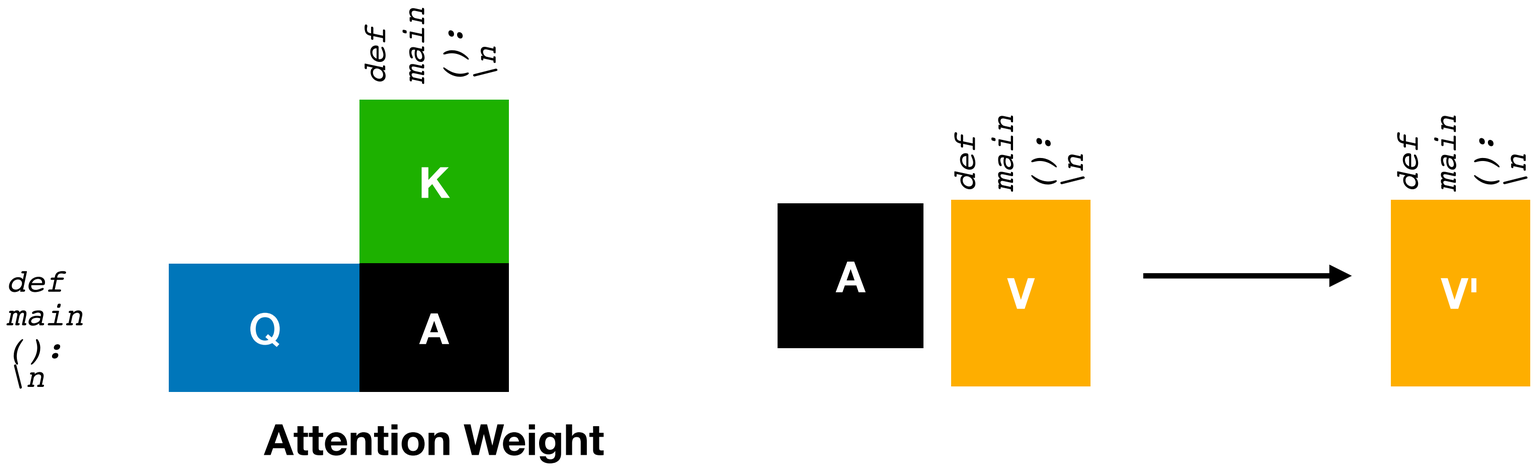
Remove null-space component



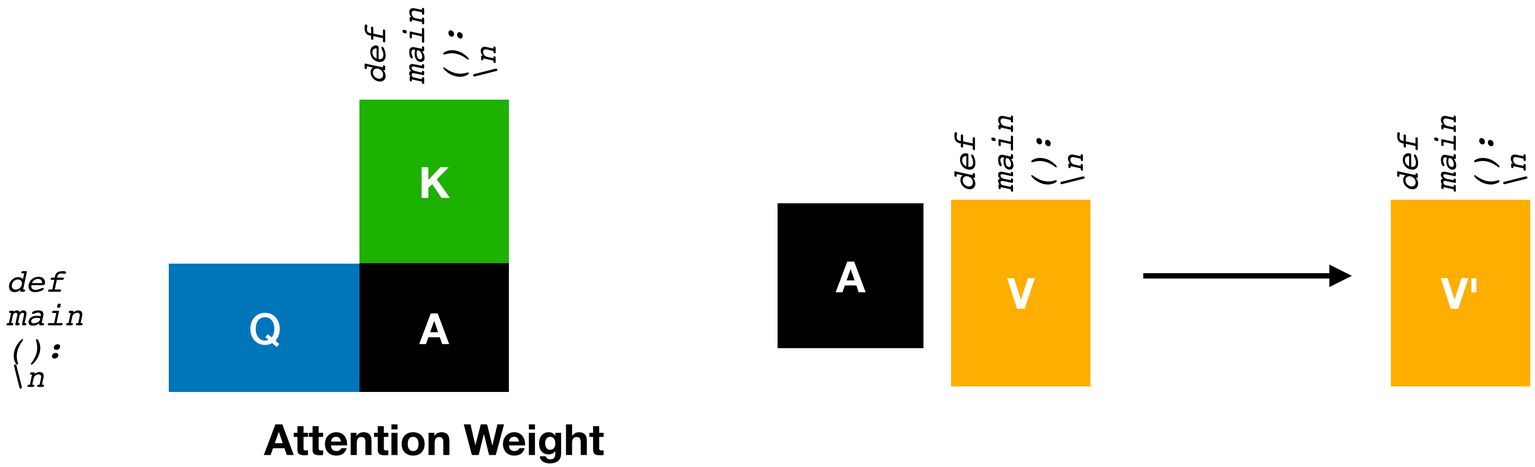
Remove null-space component



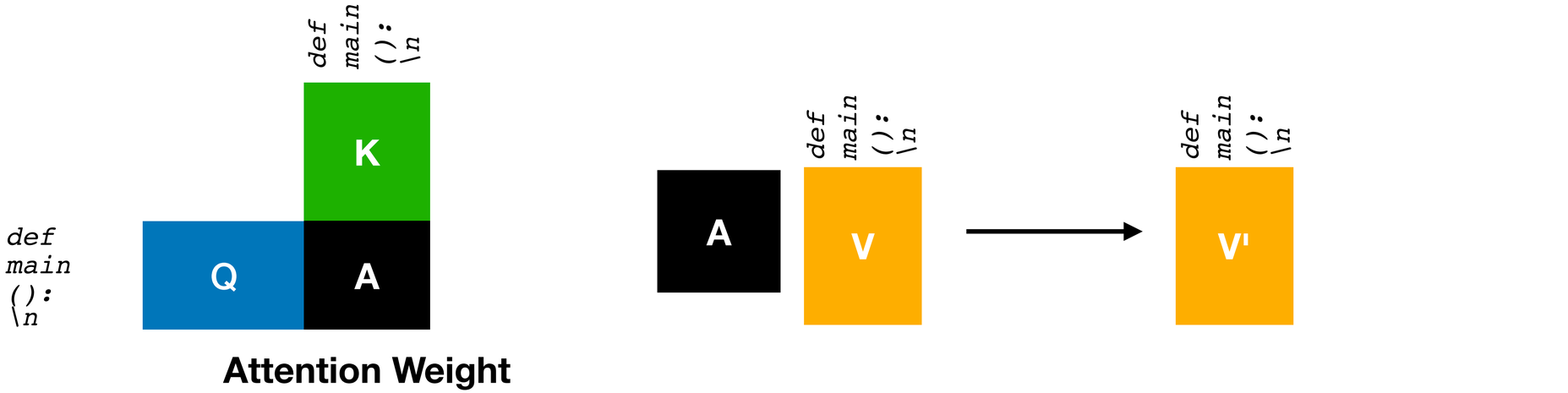
Remove null-space component



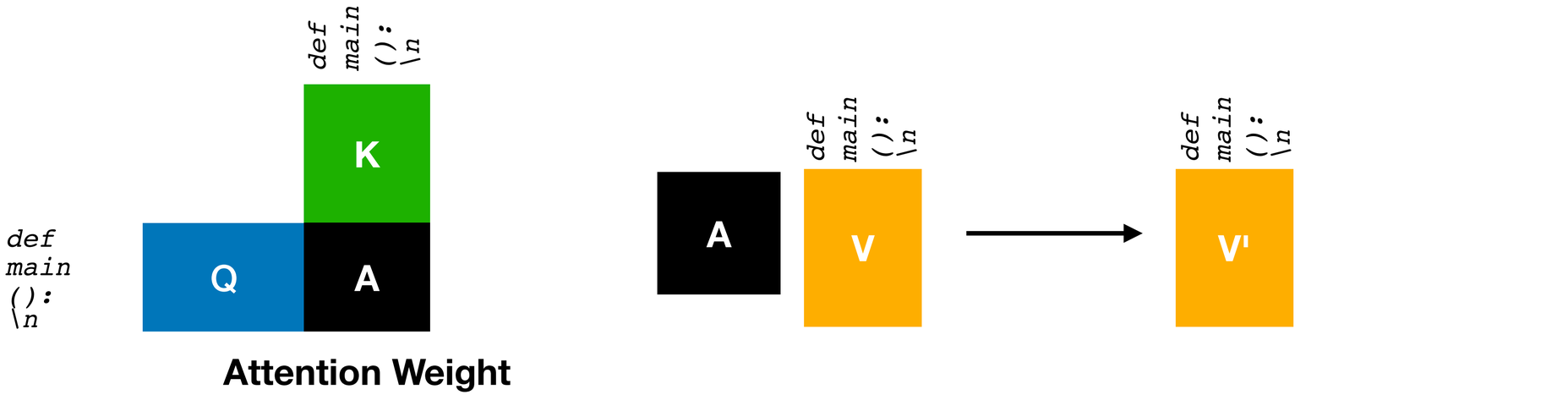
Remove null-space component



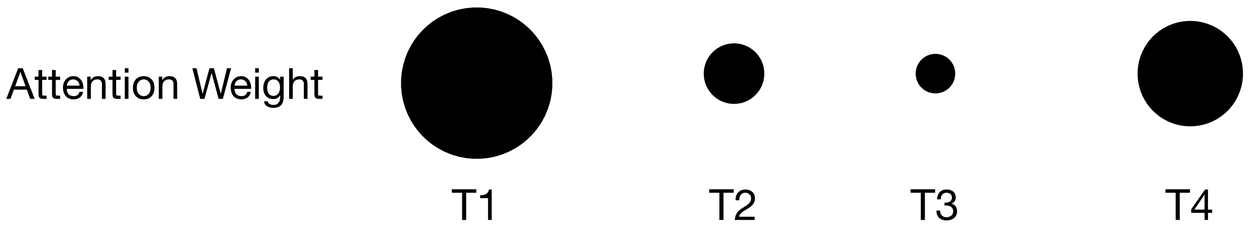
Remove null-space component



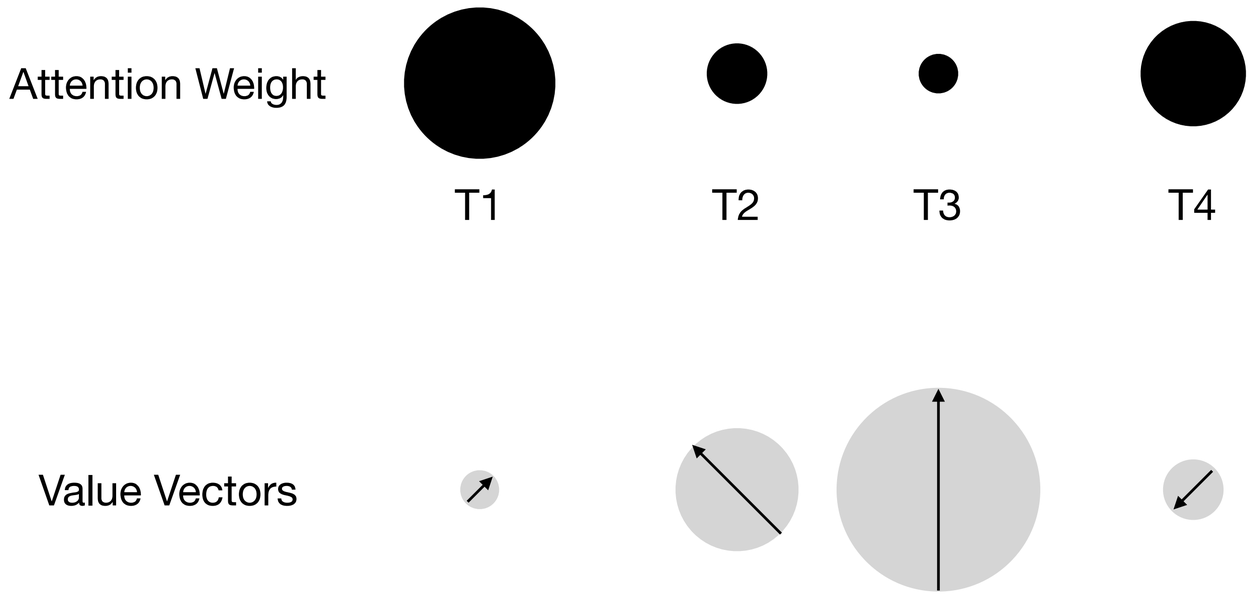
Remove null-space component



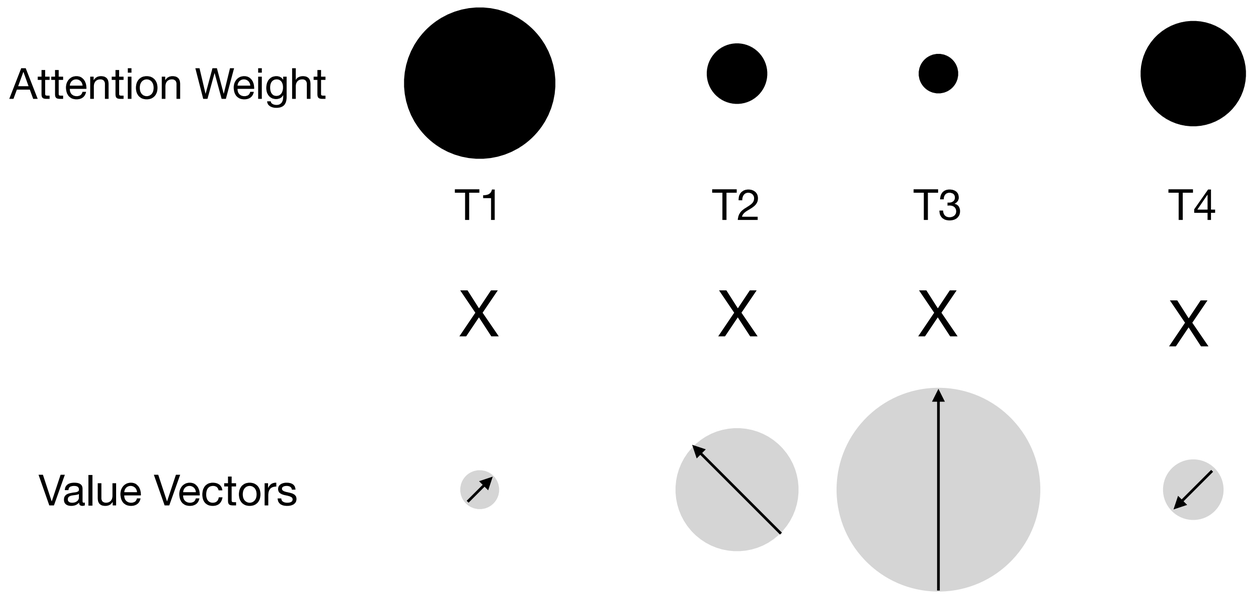
Consider size of value vector



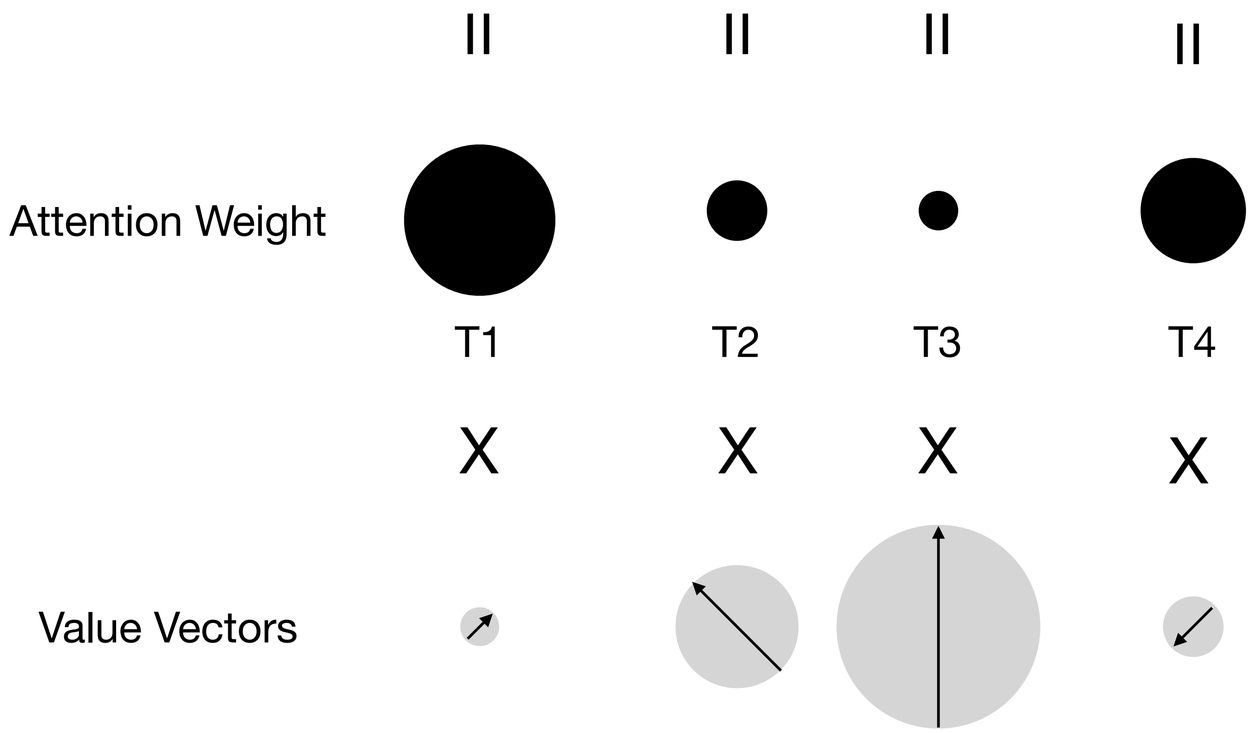
Consider size of value vector



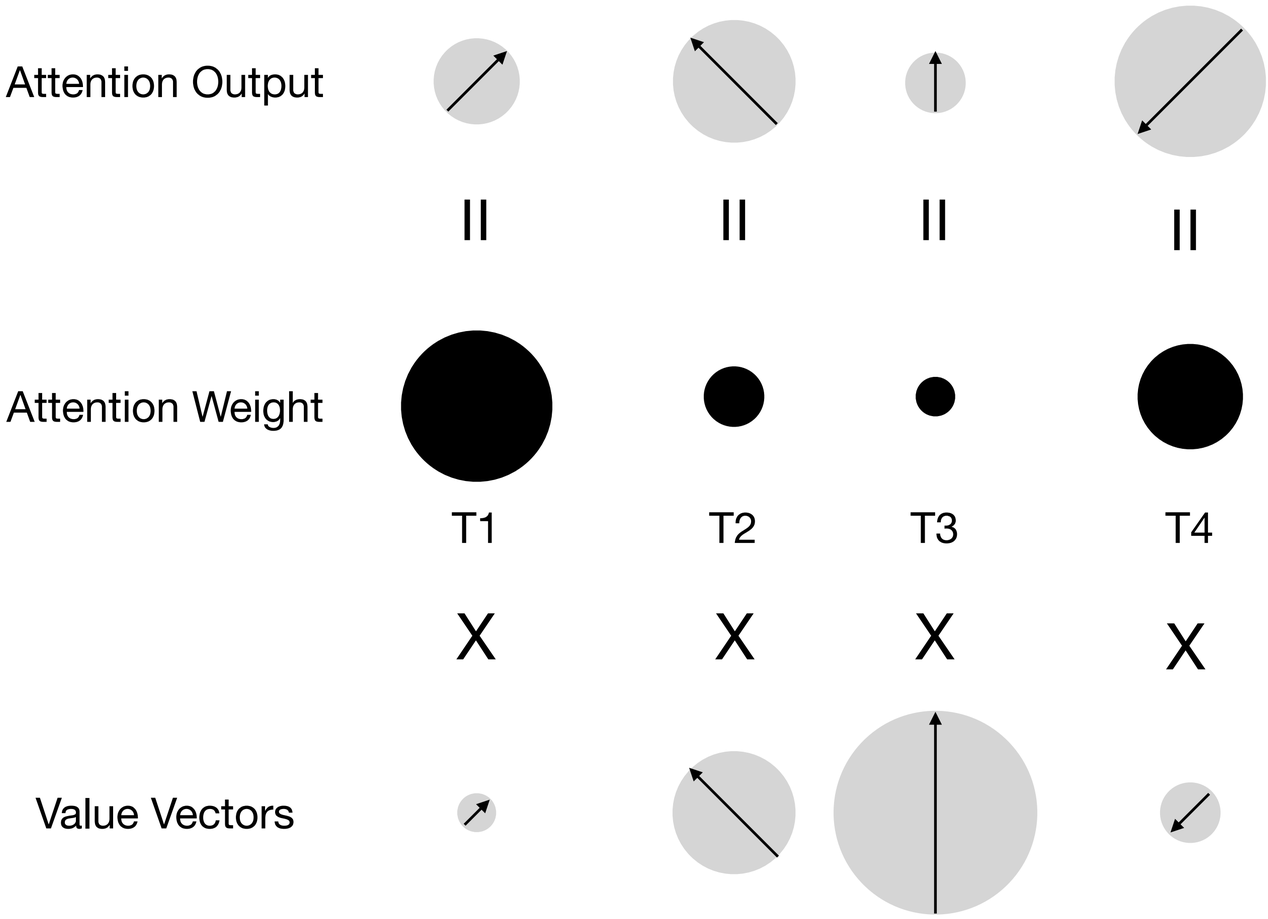
Consider size of value vector



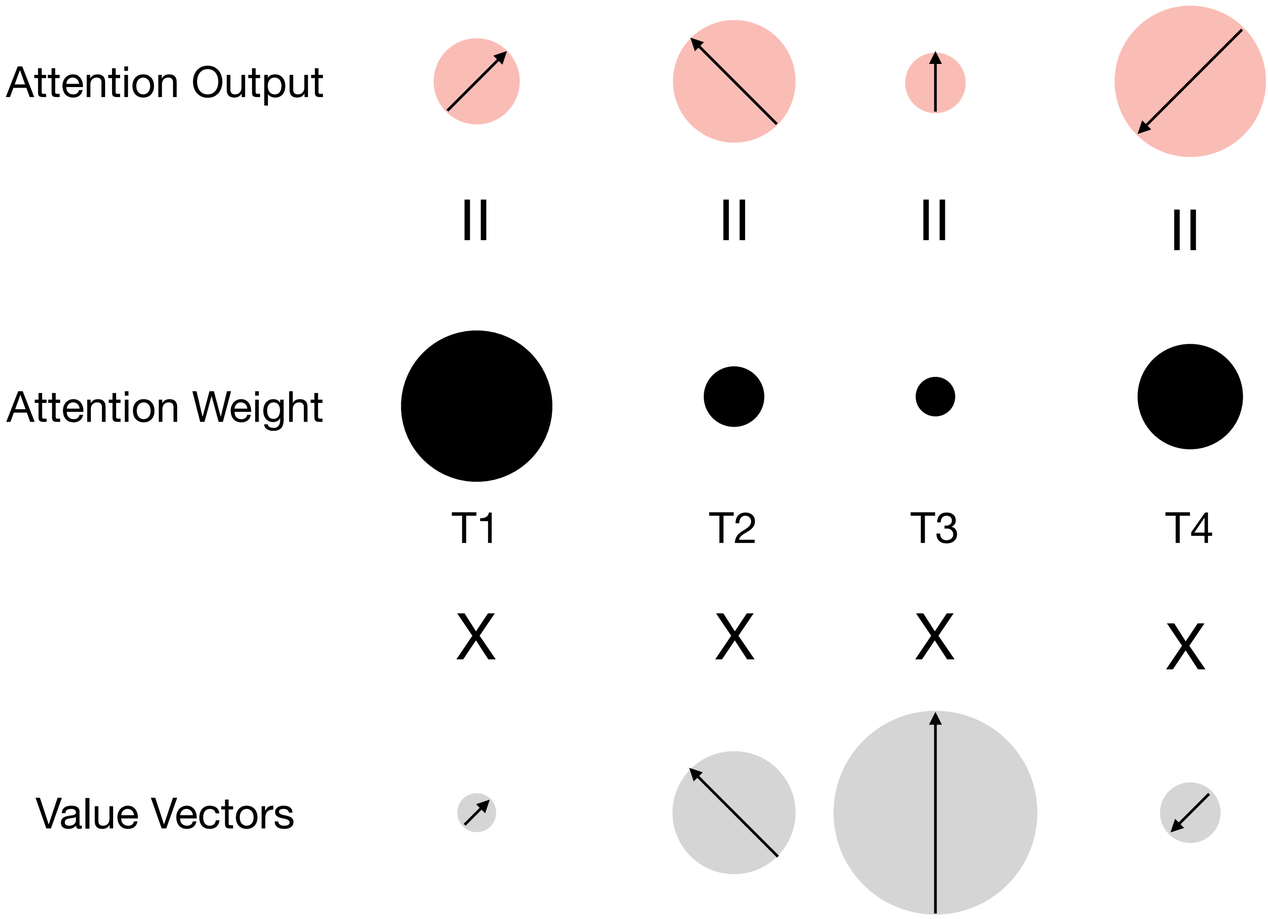
Consider size of value vector



Consider size of value vector



Consider size of value vector



Evaluation

- Qualitative Evaluation

Evaluation

- Qualitative Evaluation
  - Easy to use?

Evaluation

- Qualitative Evaluation
  - Easy to use?
  - Interactive?

Evaluation

- Qualitative Evaluation
  - Easy to use?
  - Interactive?
  - Intuitive?

Evaluation

- Qualitative Evaluation
  - Easy to use?
  - Interactive?
  - Intuitive?
  - Suitable for code?

Evaluation

- Qualitative Evaluation
  - Easy to use?
  - Interactive?
  - Intuitive?
  - Suitable for code?
- Quantitative Evaluation

Evaluation

- Qualitative Evaluation
  - Easy to use?
  - Interactive?
  - Intuitive?
  - Suitable for code?
- Quantitative Evaluation
  - Rendering speed

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X
BertViz	O	O	X	X

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X
BertViz	O	O	X	X
CircuitsViz	X	O	O	O

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X
BertViz	O	O	X	X
CircuitsViz	X	O	O	O
AttentionViz	X	O	O	X

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X
BertViz	O	O	X	X
CircuitsViz	X	O	O	O
AttentionViz	X	O	O	X
Atten-Scope	O	O	O	O

Qualitative Evaluation

	Usability	Interactive	Intuitive	Suitability
Naive Heatmap	X	X	X	X
BertViz	O	O	X	X
CircuitsViz	X	O	O	O
AttentionViz	X	O	O	X
Atten-Scope	O	O	O	O

Qualitative Evaluation

- Average Rendering Speed

Atten-Scope	CircuitsViz
389.7ms	9,275ms

Qualitative Evaluation

- Average Rendering Speed

Atten-Scope	CircuitsViz
389.7ms	9,275ms

96% Decreased

Future work

- Accumulative Attention

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer
  - I expect we can get more intuition from the more compressed info

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer
  - I expect we can get more intuition from the more compressed info
- Comparison View

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer
  - I expect we can get more intuition from the more compressed info
- Comparison View
  - A view that compares two types of attention

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer
  - I expect we can get more intuition from the more compressed info
- Comparison View
  - A view that compares two types of attention
- Information flow view

Future work

- Accumulative Attention
  - Accumulative attention is a view that combines the attention of each layer
  - I expect we can get more intuition from the more compressed info
- Comparison View
  - A view that compares two types of attention
- Information flow view
  - A view that shows information flow of each token

Summary

- Atten-Scope is a new tool for **interpreting language model**
- Atten-Scope visualize refined attention weight
- Refinement techniques
  - Remove null component
  - Consider value vector
- Atten-Scope **outperforms** compared with previous tools
- https://github.com/duncan020313/Atten-Scope.git
- https://github.com/duncan020313/Atten-Scope-Backend