On Improving Trace Analysis With System Call Arguments

Quentin Fournier

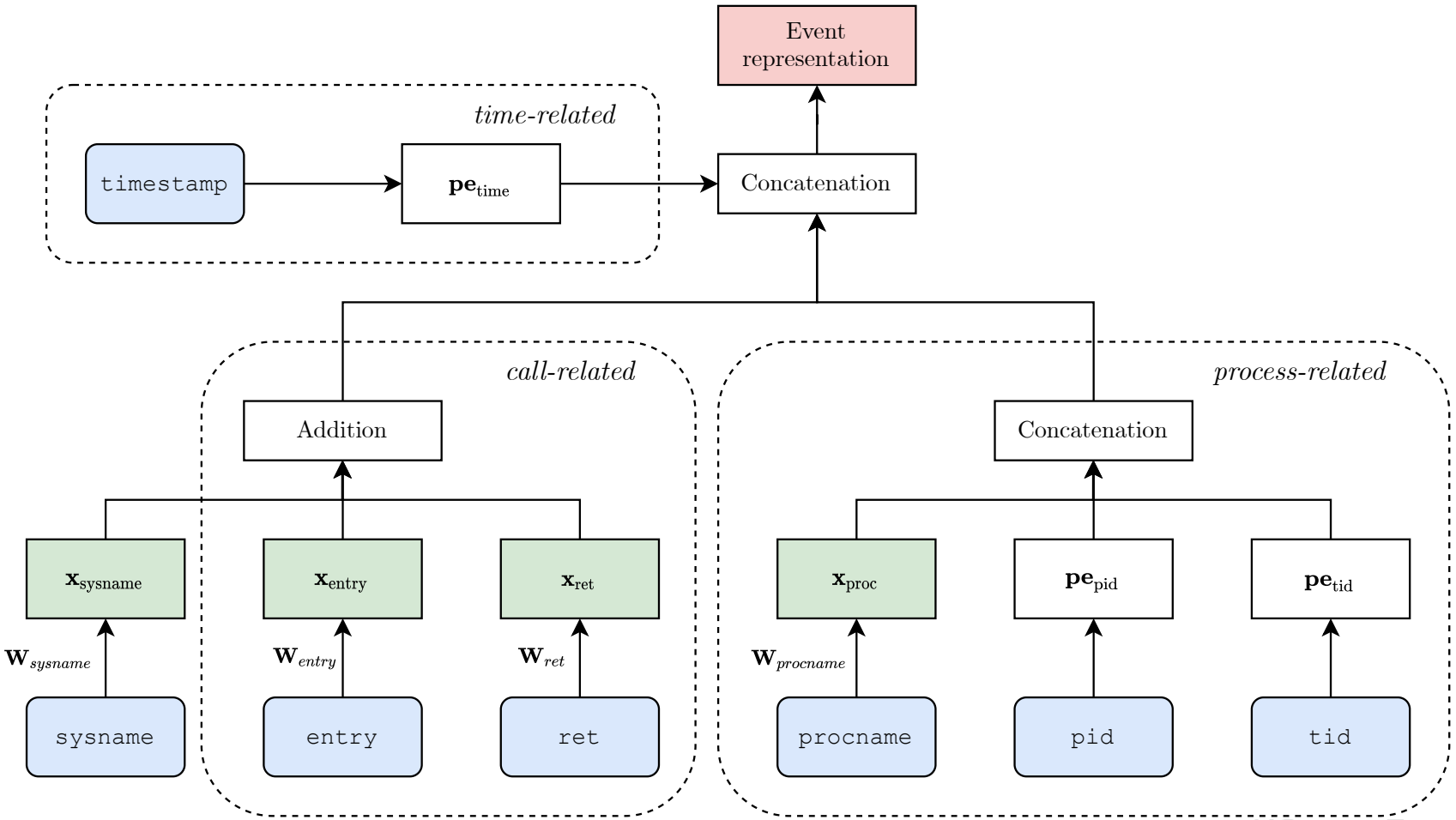
Polytechnique de Montréal
Laboratoire DORSAL

Why Use the Event Fields?

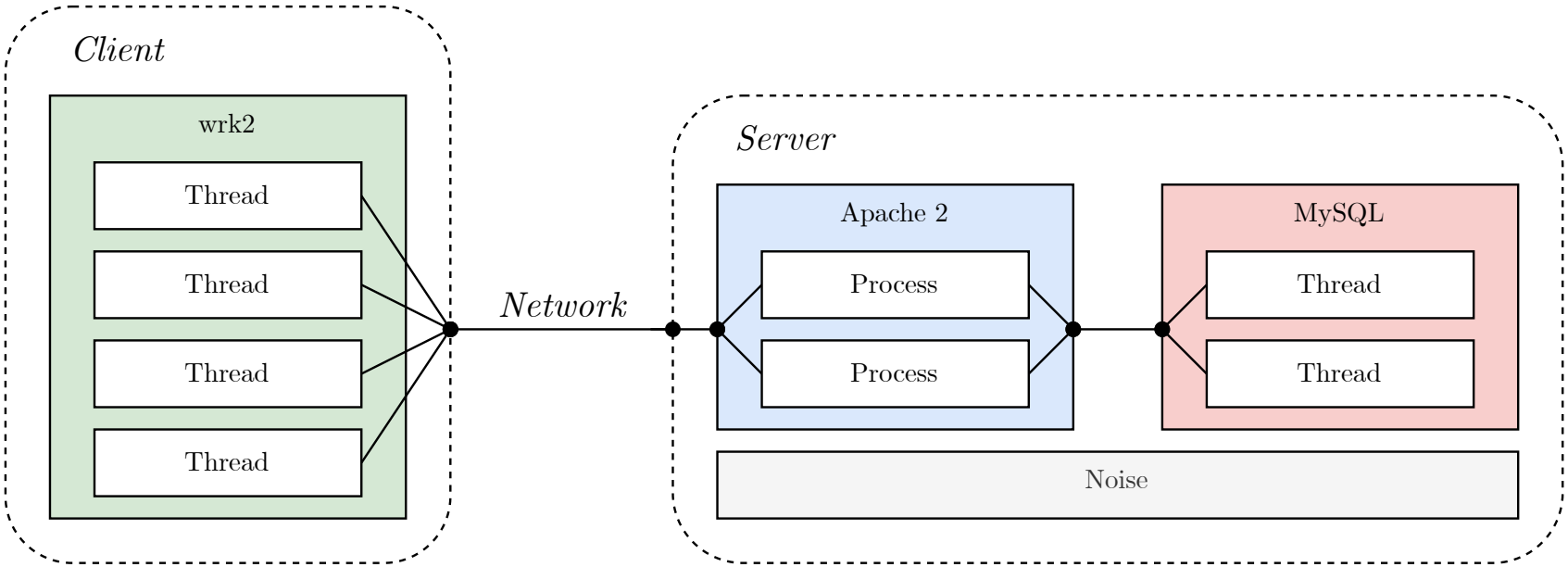
- The information is already available...
- ...or it can be made available at a low cost.
- The more information is considered, the better will be the model[1].
- In particular neural networks strive with more data.

[1] There are some caveats, for example, models may be subject to the curse of dimensionality.

How to Leverage the Arguments?

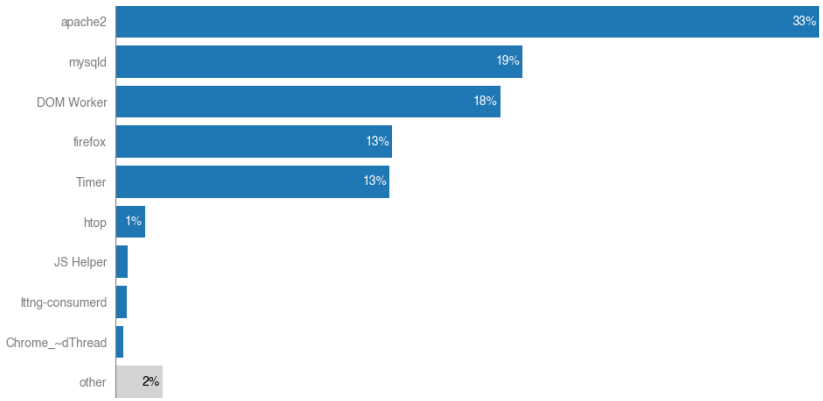


Datasets - Web Requests

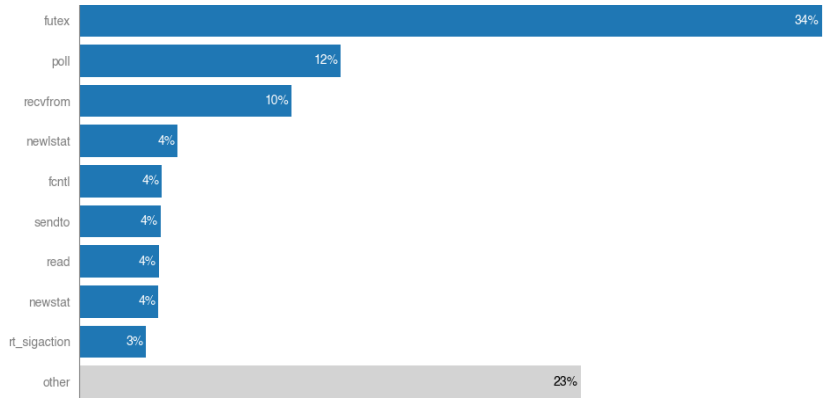


Datasets - Web Requests

- 500,000+ sequences
- 256 system calls



(a) Distribution of process names.



(b) Distribution of system call names.

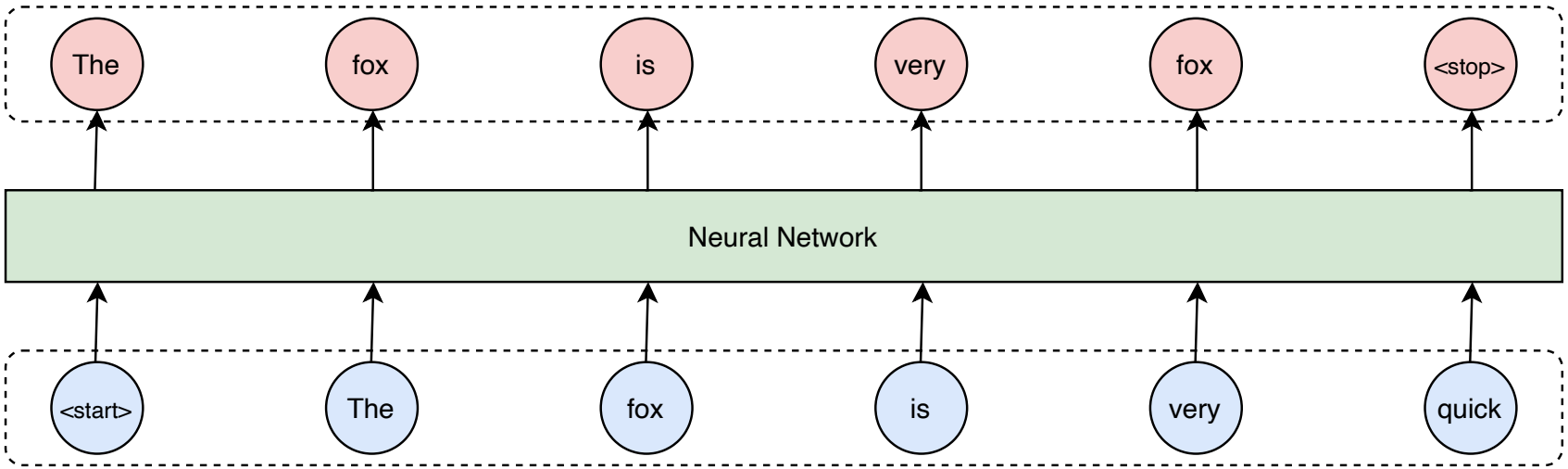
Datasets - Ciena

- Collected on pre-production servers
- 250,000+ sequences
- 256 system calls

(c) Distribution of process names.

(d) Distribution of system call names.

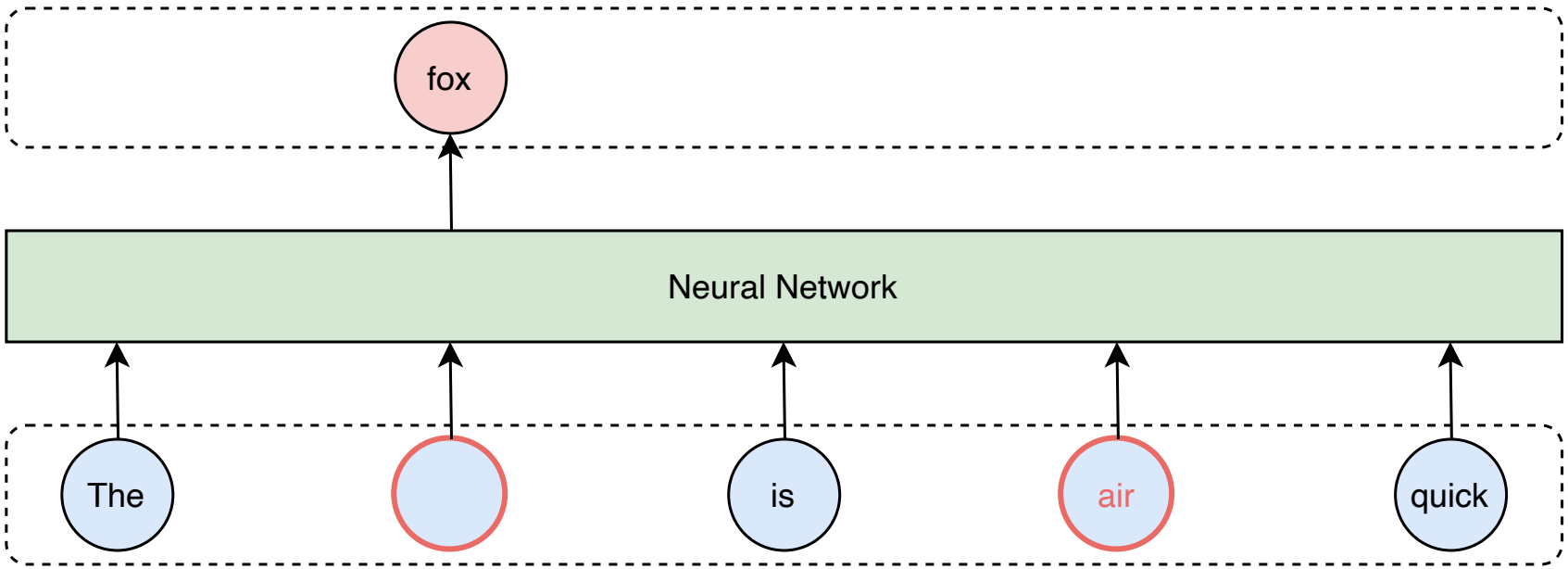
Objectives - Left-to-Right Language Model (LM)



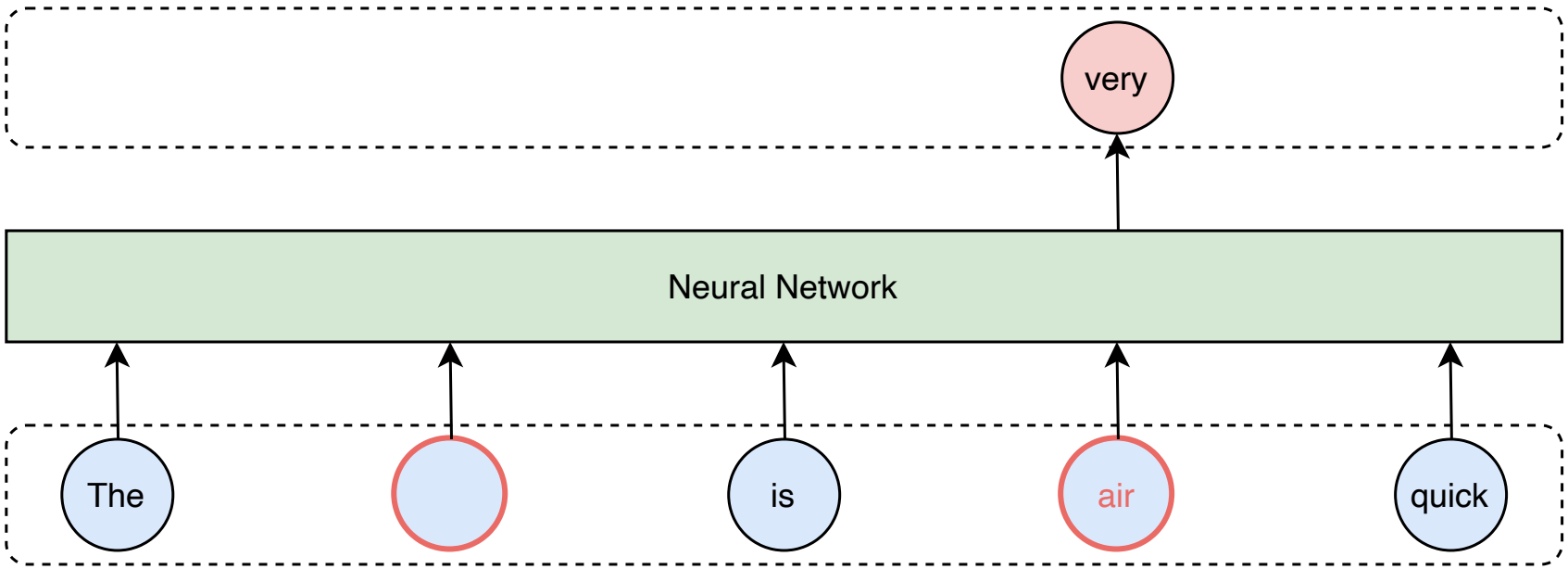
Objectives - Left-to-Right Language Model (LM)

- Computationally efficient
- Allows computing the sequence likelihood
- Could detect changes in behaviour and anomalies (future work)

Objectives - Masked Language Model (MLM)



Objectives - Masked Language Model (MLM)



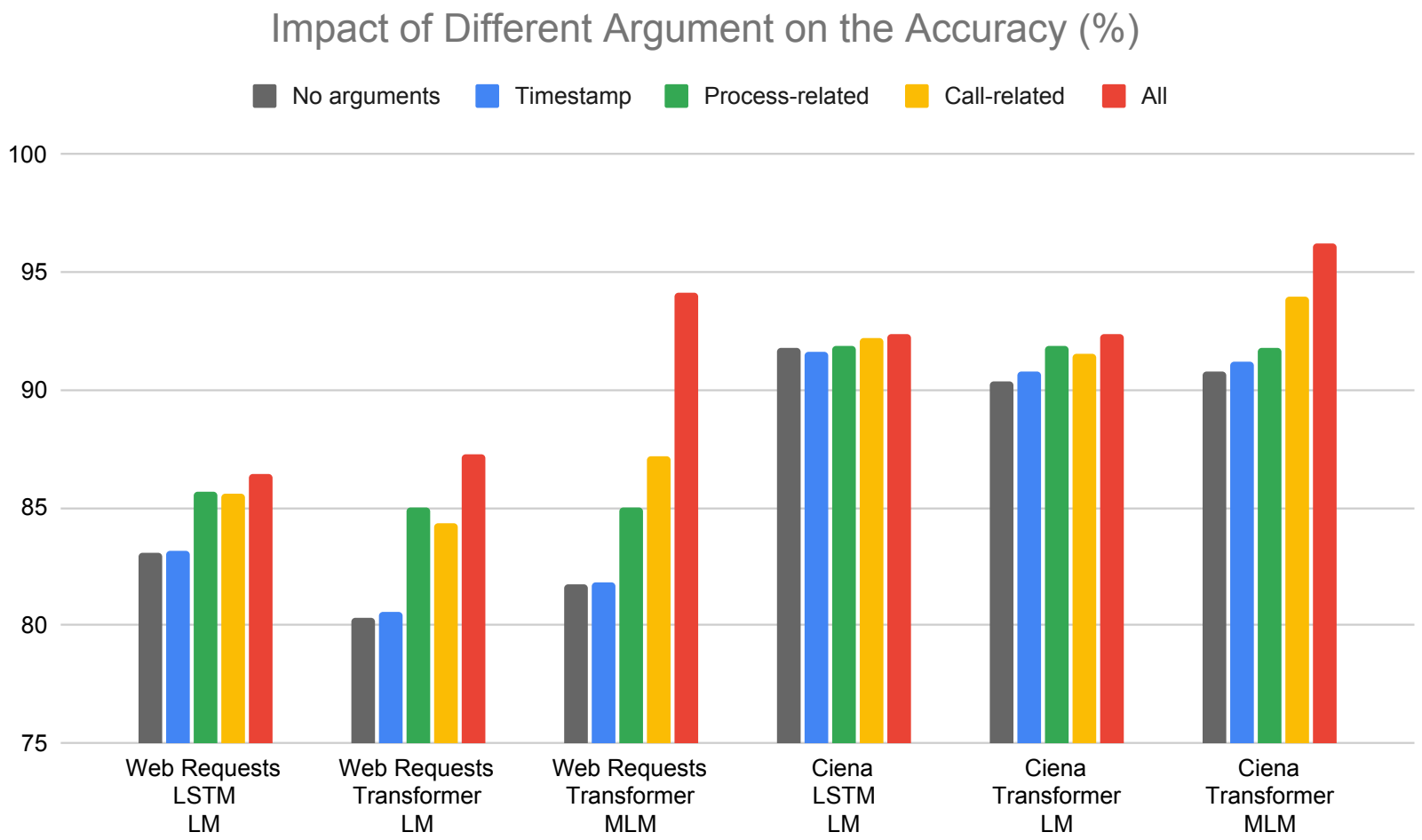
Objectives - Masked Language Model (MLM)

- Used to pre-train models
- Improve performance on downstream task
- Computationally inefficient

Neural Networks

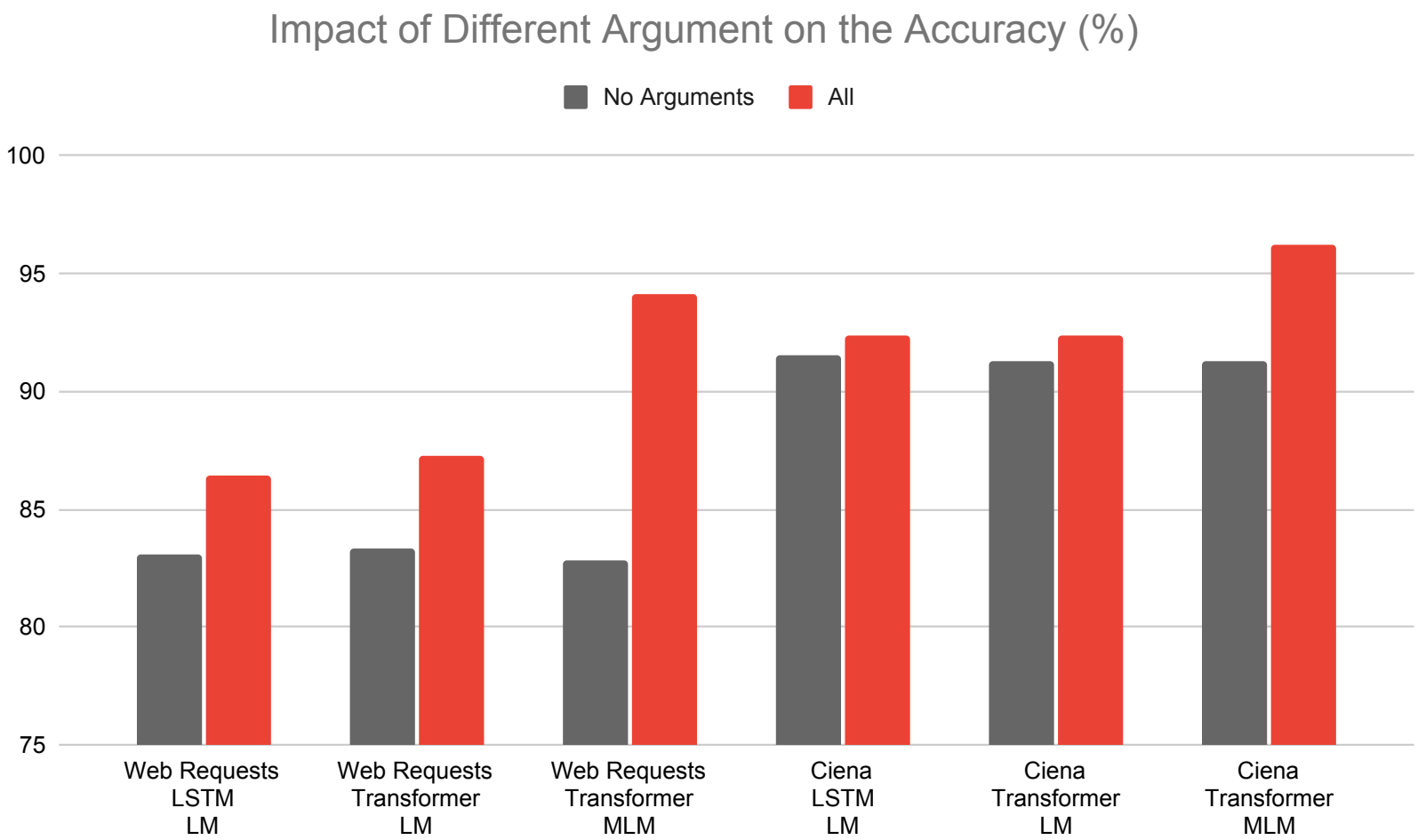
- Long-Short Term Memory (LSTM)
  - Widely popular
  - Efficient for small sequences
- Transformer
  - Current state-of-the-art
  - Relate any two events in a sequence
  - Quadratic complexity $O(N^2)$

Results


Impact of Different Argument on the Accuracy (%)
No arguments
Timestamp
Process-related
Call-related
All
100
95
90
85
80
75
Web Requests LSTM LM
Web Requests Transformer LM
Web Requests Transformer MLM
Ciena LSTM LM
Ciena Transformer LM
Ciena Transformer MLM

Ablation Study

Are models better only because they are bigger?



Future Work

- Compute the likelihood of sequences → Language model
- Detect anomalies → State-of-the-art network
- Process very long sequences → Efficient network
- Adapt to changes in behaviour → Online learning

Thank You