



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON



Multi Resolution Analysis (MRA) for Approximate Self-Attention

Zhanpeng Zeng



Sourav Pal



Jeffery Kline



Glenn Fung



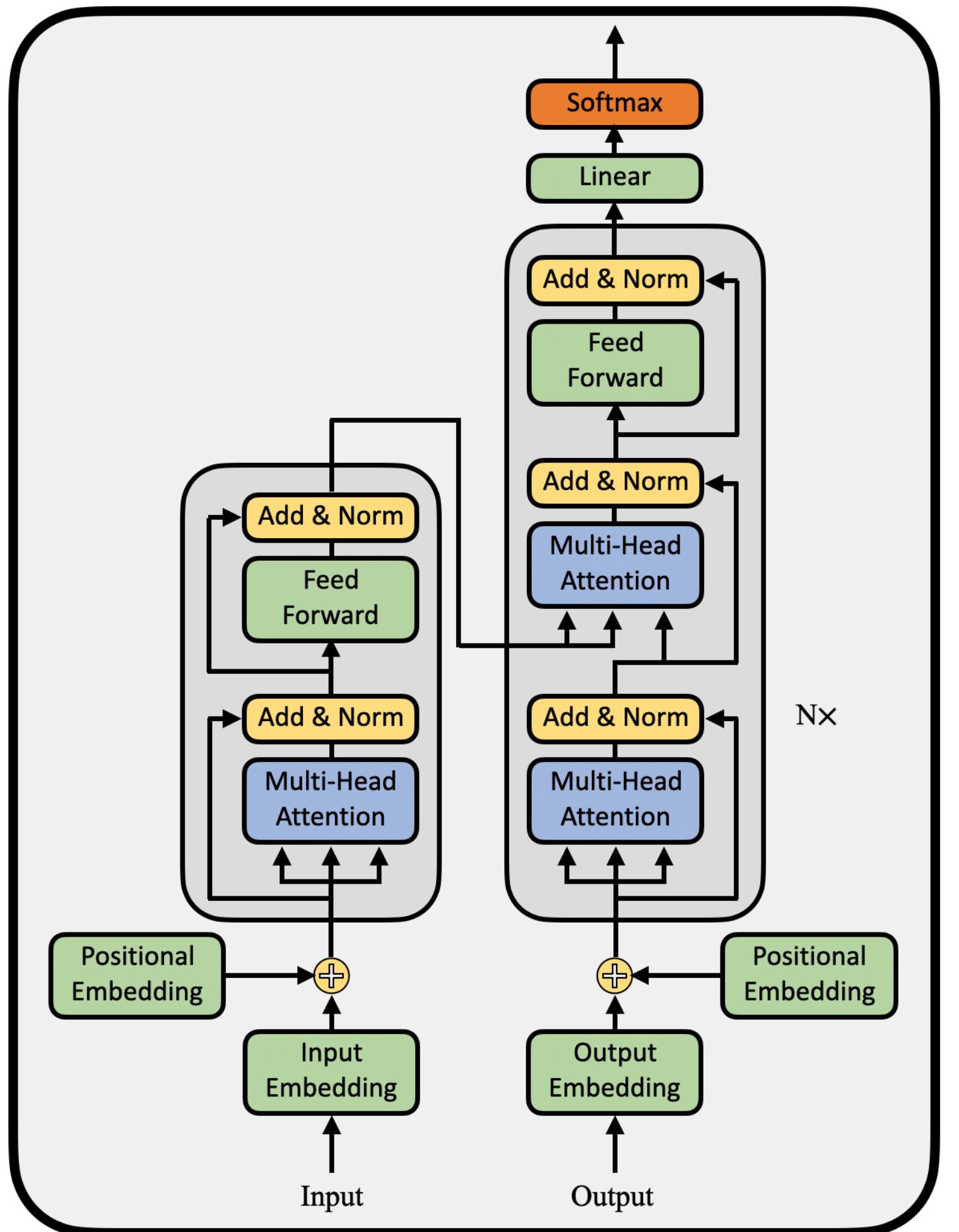
Vikas Singh



Motivation Transformer

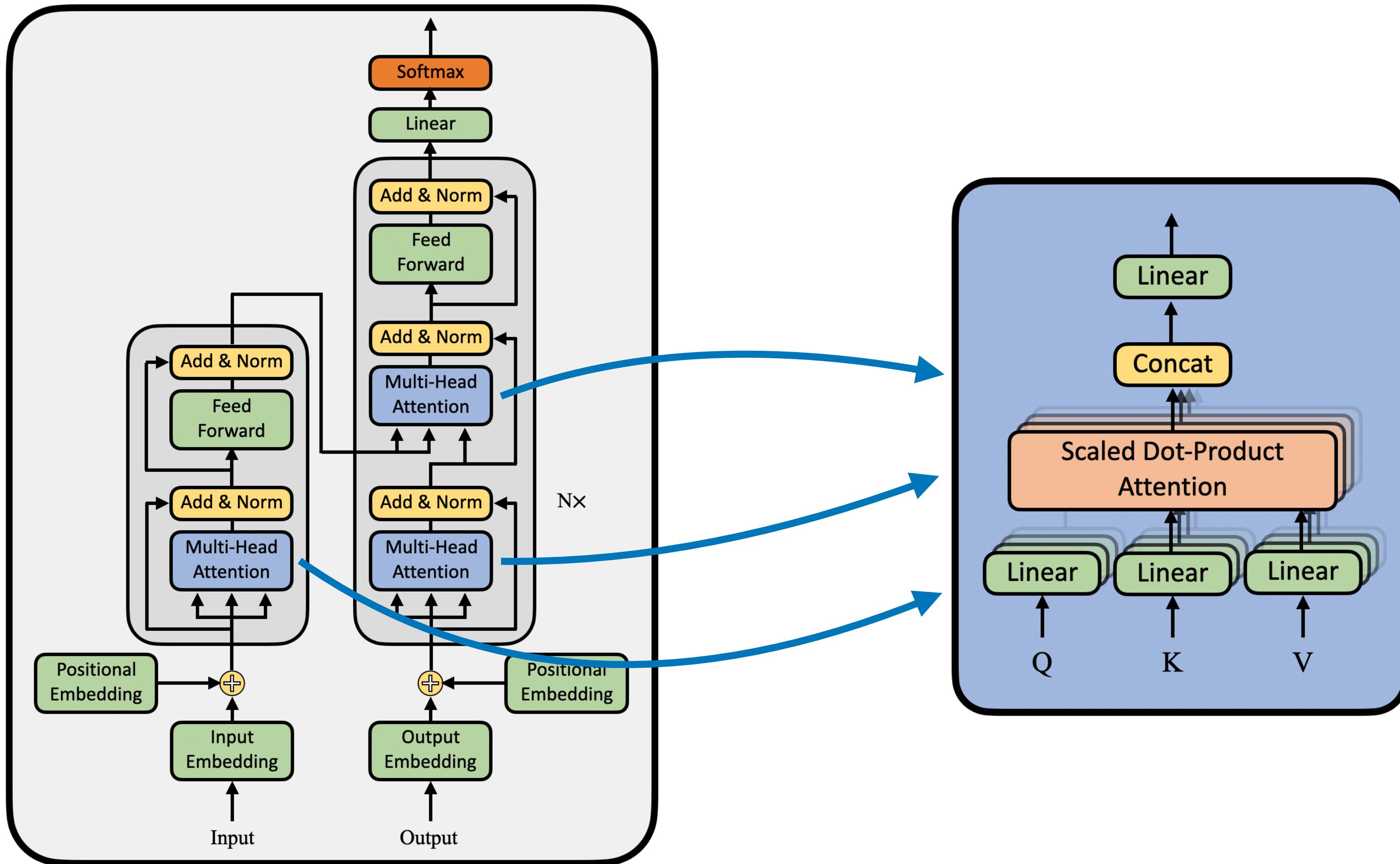
Motivation

Transformer



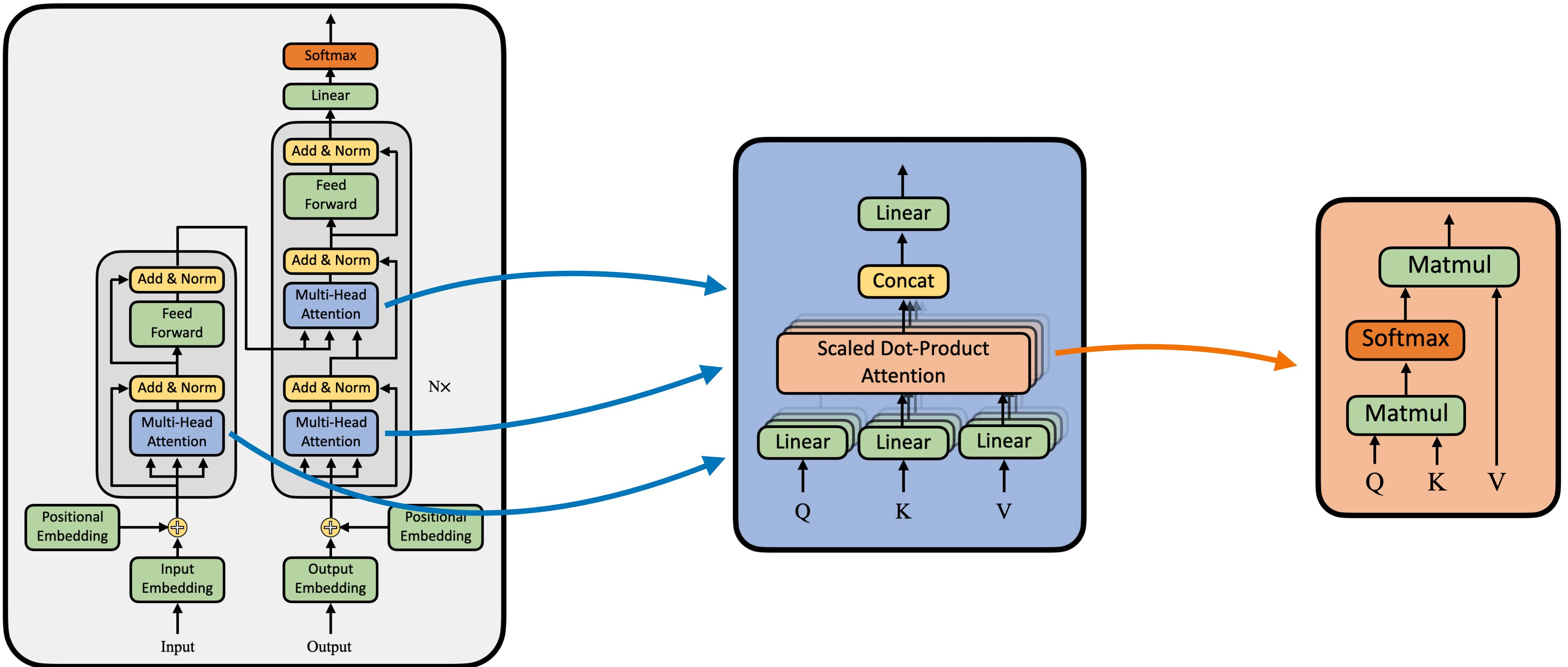
Motivation

Transformer



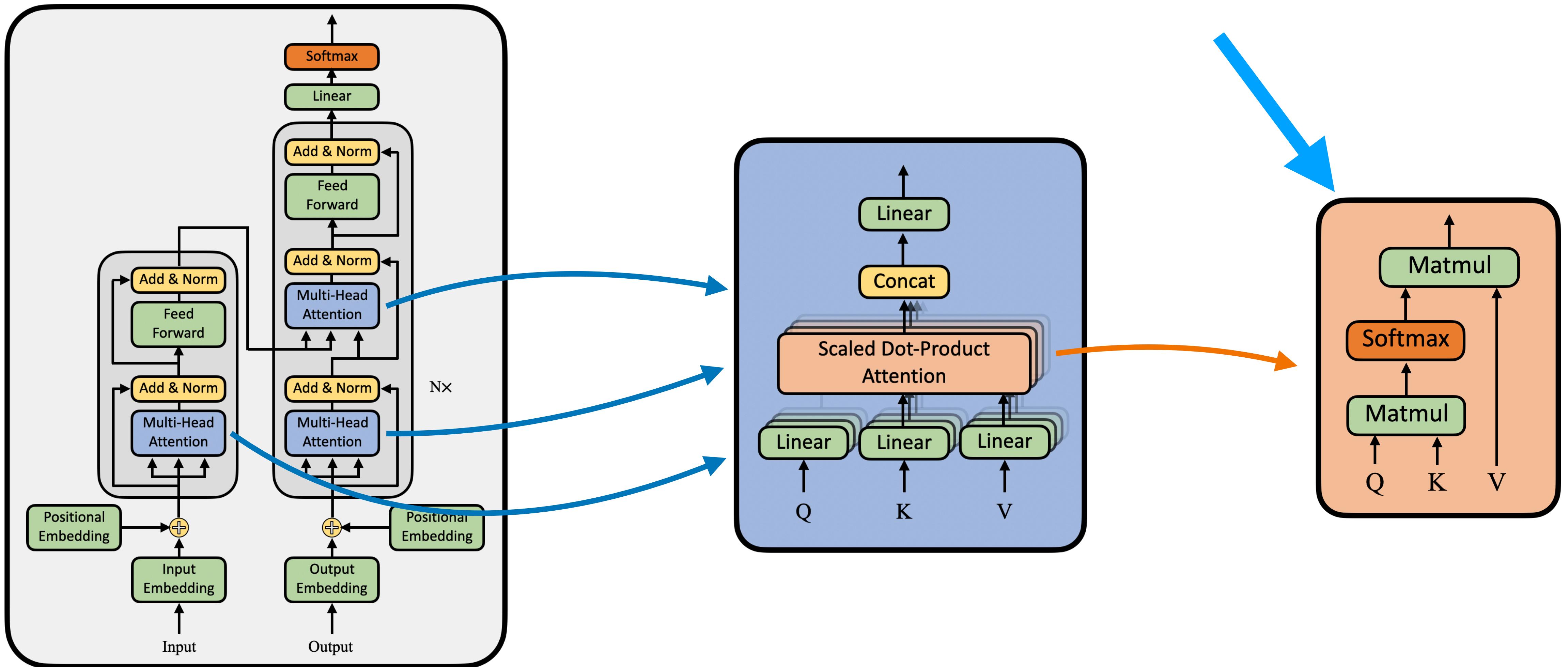
Motivation

Transformer



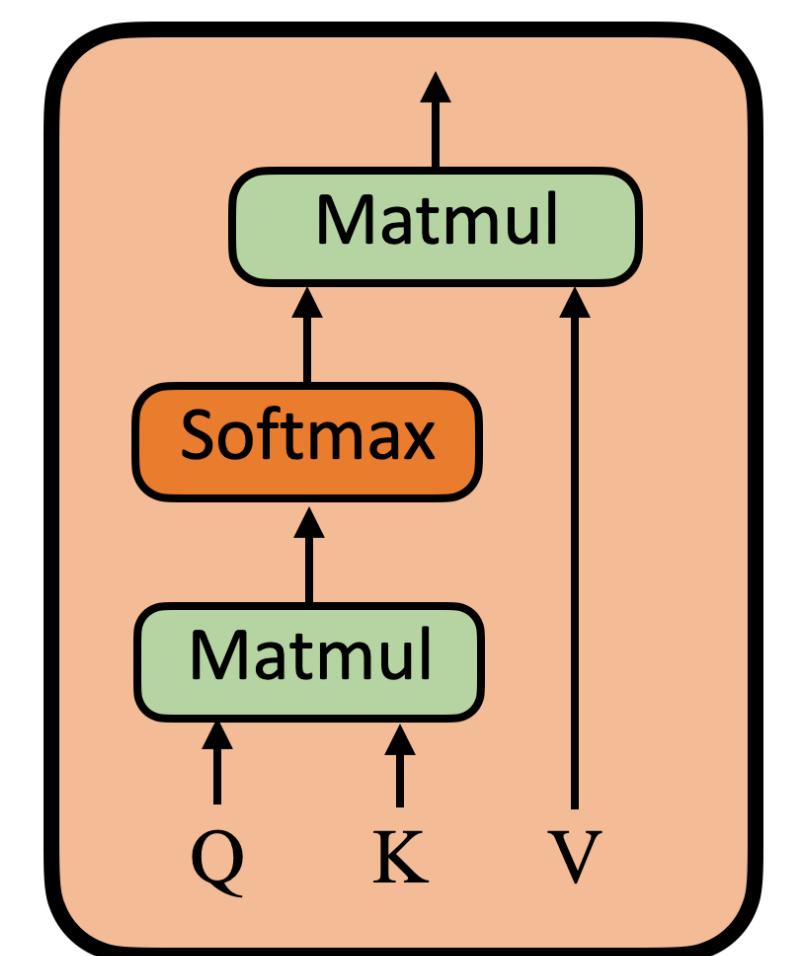
Motivation

Transformer



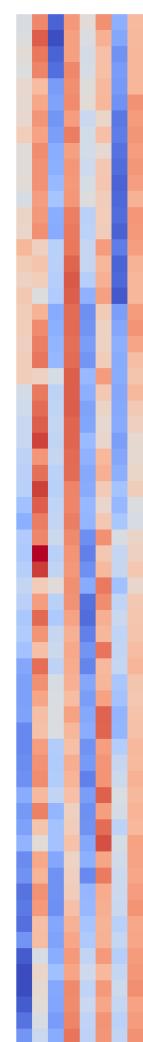
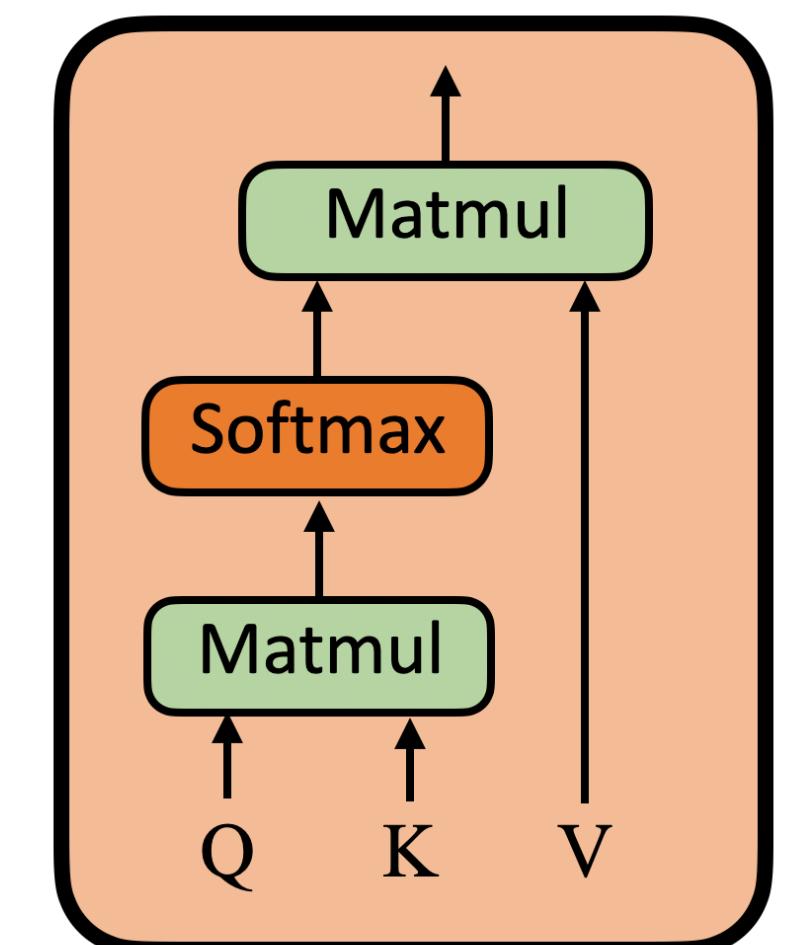
Motivation

Transformer

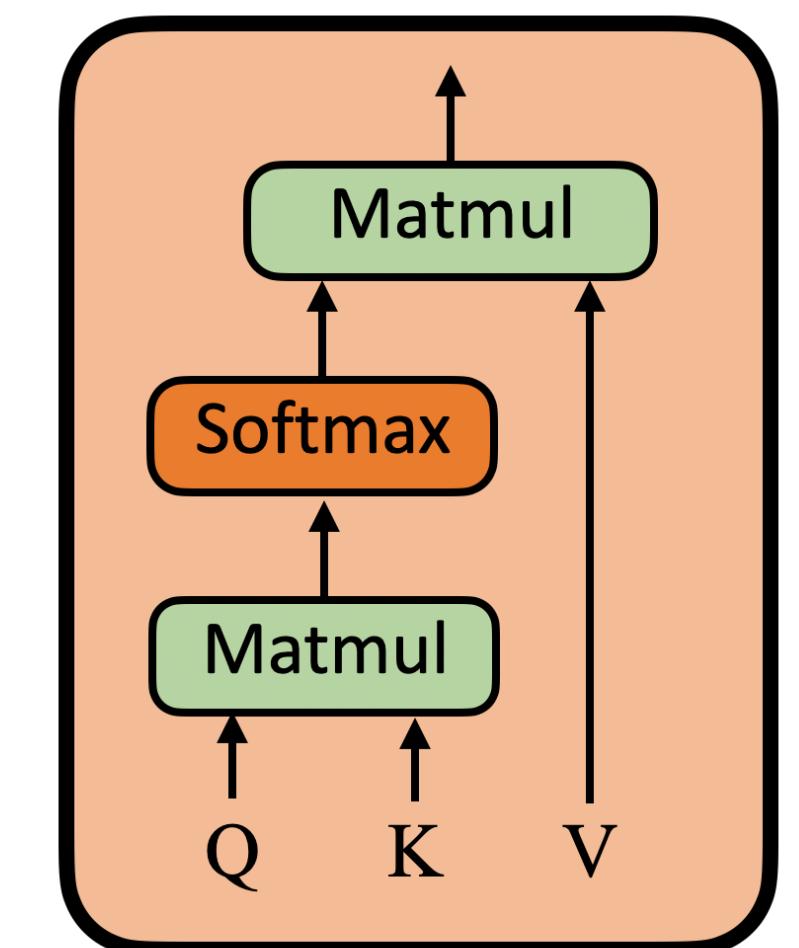
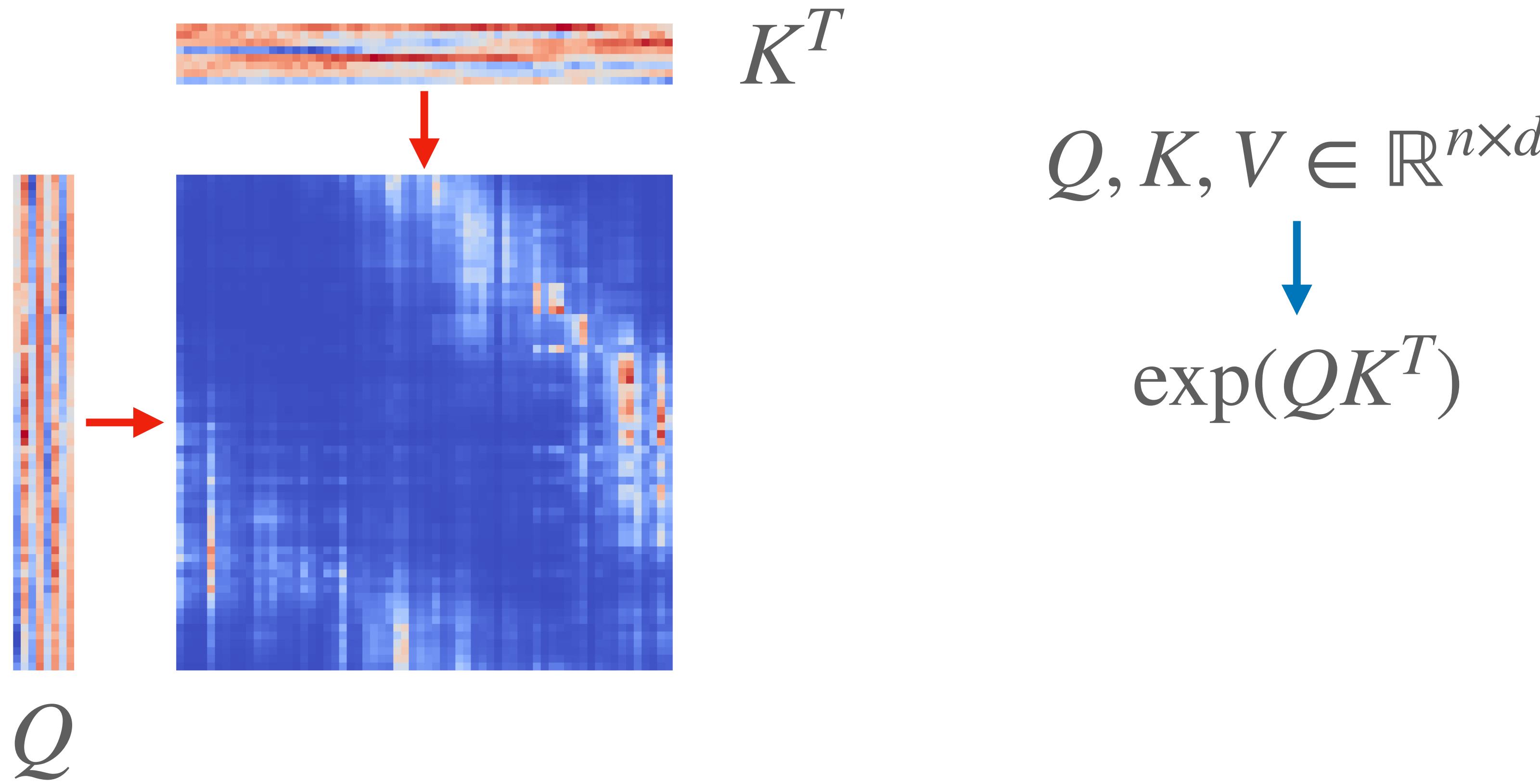


Motivation

Transformer

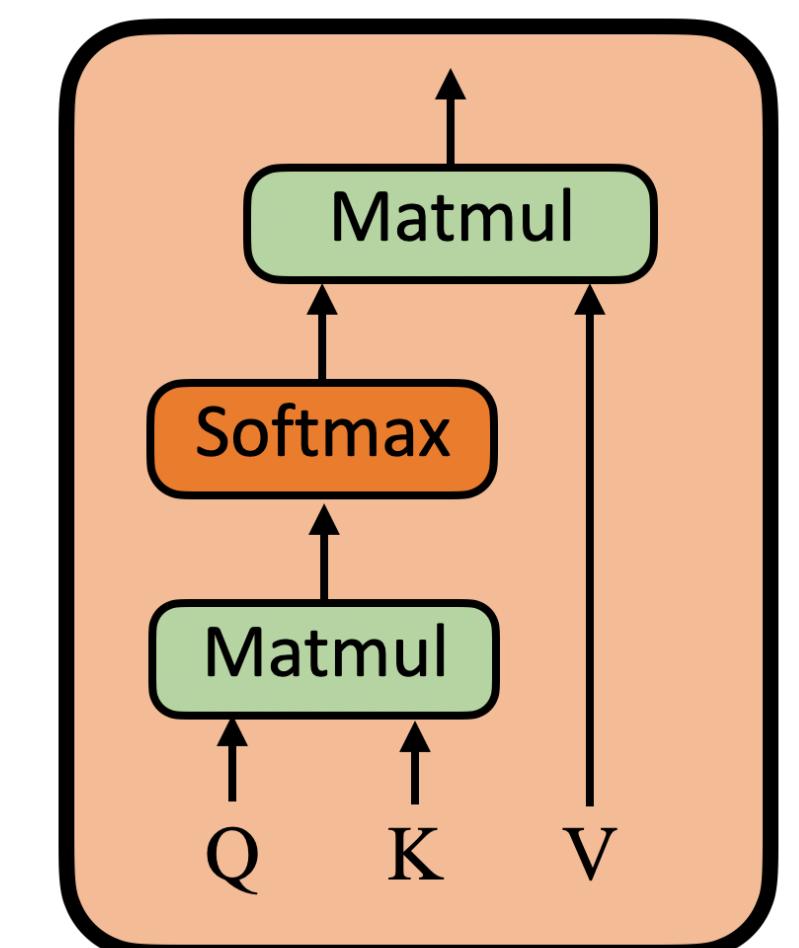
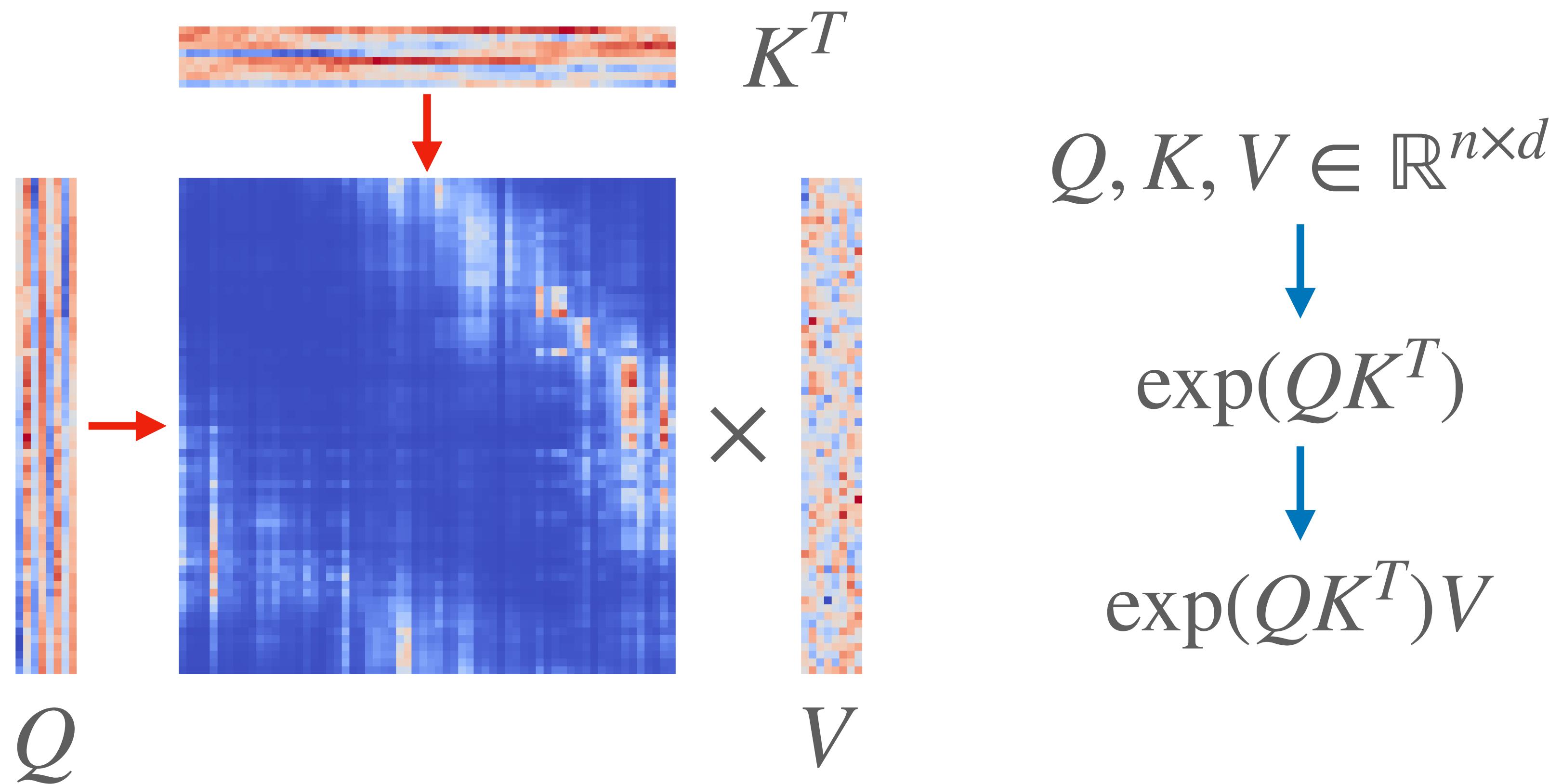
 Q  K^T $Q, K, V \in \mathbb{R}^{n \times d}$ 

Motivation Transformer

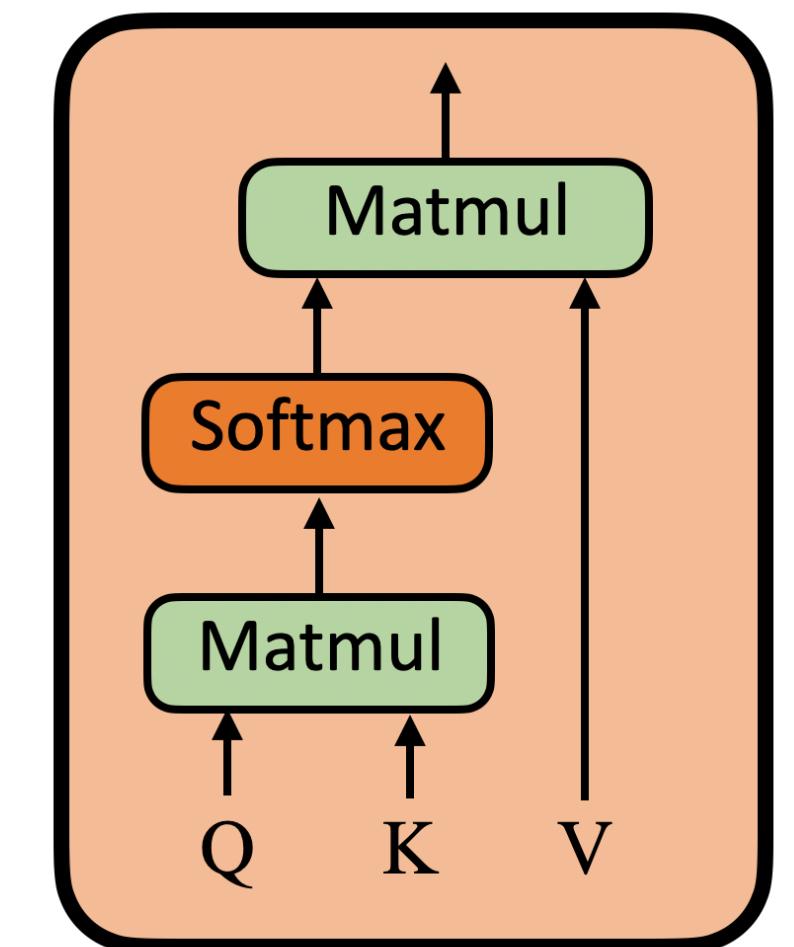
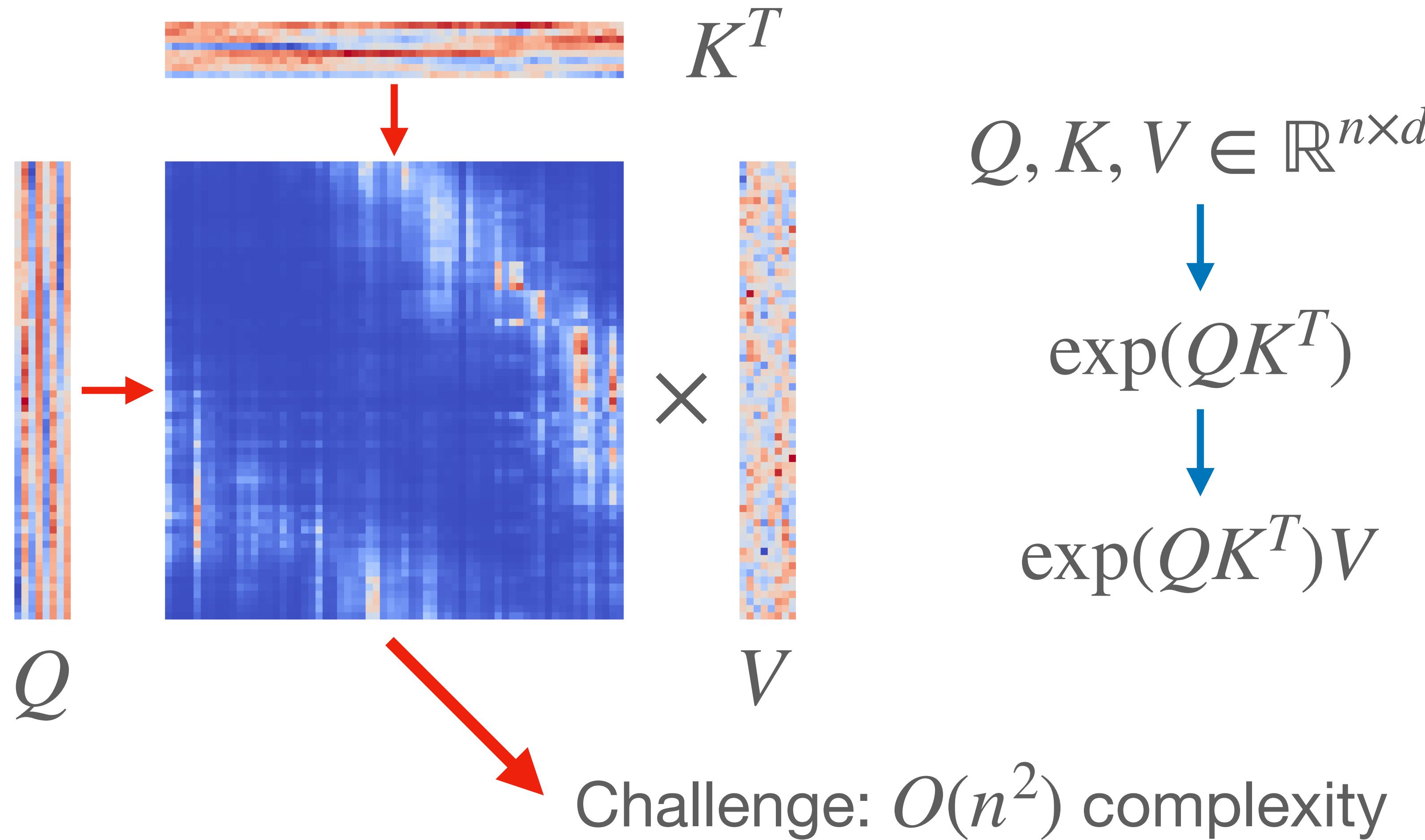


Motivation

Transformer



Motivation Transformer



Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev*
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

Lukasz Kaiser*
Google Research
{lukaszkaiser,levsk}

Longformer: The Long-Document Transformer

H-Transformer-1D: Fast One-Dimensional Hierarchical Attention for Sequences

Zhenhai Zhu
Google Research
zhenhai@google.com

Radu Soricut
Google Research
rsoricut@google.com

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Linformer: Self-Attention with Linear Complexity

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, Hao Ma
Facebook AI, Seattle, WA
{sinongwang, belindali, hanfang, mkhabsa, haom}@fb.com

Longformer: The Long-Document Transformer

Fast One-Dimensional Hierarchical Attention for Sequences

-free Transformer with Linear Complexity

Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention

Yunyang Xiong¹ **Zhanpeng Zeng**¹ **Rudrasis Chakraborty**² **Mingxing Tan**³
Glenn Fung⁴ **Yin Li**¹ **Vikas Singh**¹

¹ University of Wisconsin-Madison ² UC Berkeley ³ Google Brain ⁴ American Family Insurance
yxiong43@wisc.edu, zzeng38@wisc.edu, rudra@berkeley.edu, tanmingxing@google.com, gfung@amfam.com,
yin.li@wisc.edu, vsingh@biostat.wisc.edu

Hang Xu³
^{1*}
Ark Lab

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Lin

Scatterbrain: Unifying Sparse and Low-rank Attention Approximation

Beidi Chen^{*†}, Tri Dao^{*†}, Eric Winsor[†], Zhao Song[§], Atri Rudra[‡], and Christopher Ré[†]

[†]Department of Computer Science, Stanford University

[§]Adobe Research

Nyström

[‡]Department of Computer Science and Engineering, University at Buffalo, SUNY

{beidic, trid, winsor}@stanford.edu, zsong@adobe.com, atri@buffalo.edu,
chrismre@cs.stanford.edu

Yun

Glen

¹ University of Wisconsin-Madison
yxiong43@wisc.edu, zzeng38@wisc.edu

Zhanpeng Zeng¹ Yunyang Xiong¹ Sathya N. Ravi² Shailesh Acharya³ Glenn Fung³ Vikas Singh¹
yixiong43@wisc.edu, zzeng38@wisc.edu

Longformer: The Long-Document Transformer

Fast One-Dimensional Hierarchical Attention for

with Linear

-Attention Via Bernoulli

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev*
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

Łukasz Kaiser*
Google Research
{lukaszkaiser,levskaya}@google.com

Anselm Levskaya
Google Research
levskaya@google.com

Longformer: The Long-Document Transformer

Iz Beltagy* **Matthew E. Peters*** **Arman Cohan***
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy, matthewp, armanc}@allenai.org

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev*
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

Łukasz Kaiser*
Google Research
{lukaszkaiser, levskaya}@google.com

Anselm Levskaya
Google Research
levskaya@google.com

Longformer: The Long-Document Transformer

Iz Beltagy* Matthew E. Peters* Arman Cohan*
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy, matthewp, armanc}@allenai.org

Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

RETHINKING ATTENTION WITH PERFORMERS

Krzysztof Choromanski*, Valerii Likhoshesterov*, David Dohan*, Xingyou Song*,
Andreea Gane*, Tamas Sarlos*, Peter Hawkins*, Jared Davis*, Afroz Mohiuddin*,
Łukasz Kaiser*, David Belanger*, Lucy Colwell*, Adrian Weller*,
¹Google ²University of Cambridge ³DeepMind ⁴Alan Turing Institute

Linformer: Self-Attention with Linear Complexity

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, Hao Ma
Facebook AI, Seattle, WA
{sinongwang, belindali, hanfang, mkhabsa, haom}@fb.com

Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention

Yunyang Xiong ¹ Zhanpeng Zeng ¹ Rudrasis Chakraborty ² Mingxing Tan ³
Glenn Fung ⁴ Yin Li ¹ Vikas Singh ¹
¹ University of Wisconsin-Madison ² UC Berkeley ³ Google Brain ⁴ American Family Insurance
xiong43@wisc.edu, zzeng38@wisc.edu, rudra@berkeley.edu, tanmingxing@google.com, gfung@amfam.com,
yin.li@wisc.edu, vsingh@biostat.wisc.edu

SOFT: Softmax-free Transformer with Linear Complexity

Jiachen Lu¹ Jinghan Yao¹ Junge Zhang¹ Xiatian Zhu² Hang Xu³
Weiguo Gao¹ Chunjing Xu³ Tao Xiang² Li Zhang^{1*}
¹Fudan University ²University of Surrey ³Huawei Noah's Ark Lab

Related Works

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev*
U.C. Berkeley & Google Research
kitaev@cs.berkeley.edu

Łukasz Kaiser*
Google Research
{lukaszkaiser,levskaya}@google.com

Anselm Levskaya
Google Research
levskaya@google.com

Longform

Iz Beltagy
Allen Institute
{beltagy, iz}@allenai.org

Big Bird

Manzil Zaheer, Guru Guruganesh, Avinava Dubey,
Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham,
Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed
Google Research
{manzilz, gurug, avinavadubey}@google.com

Scatterbrain: Unifying Sparse and Low-rank Attention Approximation

Beidi Chen*†, Tri Dao*†, Eric Winsor†, Zhao Song§, Atri Rudra‡, and Christopher Ré†

†Department of Computer Science, Stanford University
§Adobe Research

‡Department of Computer Science and Engineering, University at Buffalo, SUNY
{beidic, trid, winsor}@stanford.edu, zsong@adobe.com, atri@buffalo.edu,
chrismre@cs.stanford.edu

RETHINKING ATTENTION WITH PERFORMERS

Krzysztof Choromanski*¹, Valerii Likhoshesterov*², David Dohan*¹, Xingyou Song*¹
Andreea Gane*¹, Tamas Sarlos*¹, Peter Hawkins*¹, Jared Davis*³, Afroz Mohiuddin¹
Łukasz Kaiser¹, David Belanger¹, Lucy Colwell^{1,2}, Adrian Weller^{2,4}

¹Google ²University of Cambridge ³DeepMind ⁴Alan Turing Institute

Complexity

Hao Ma
fb.com

ating Self-Attention

Mingxing Tan³
American Family Insurance
com, gfung@amfam.com,

SOFT. SOFTMAX-FREE TRANSFORMER with Linear Complexity

Jiachen Lu¹ Jinghan Yao¹ Junge Zhang¹ Xiatian Zhu² Hang Xu³
Weiguo Gao¹ Chunjing Xu³ Tao Xiang² Li Zhang^{1*}

¹Fudan University

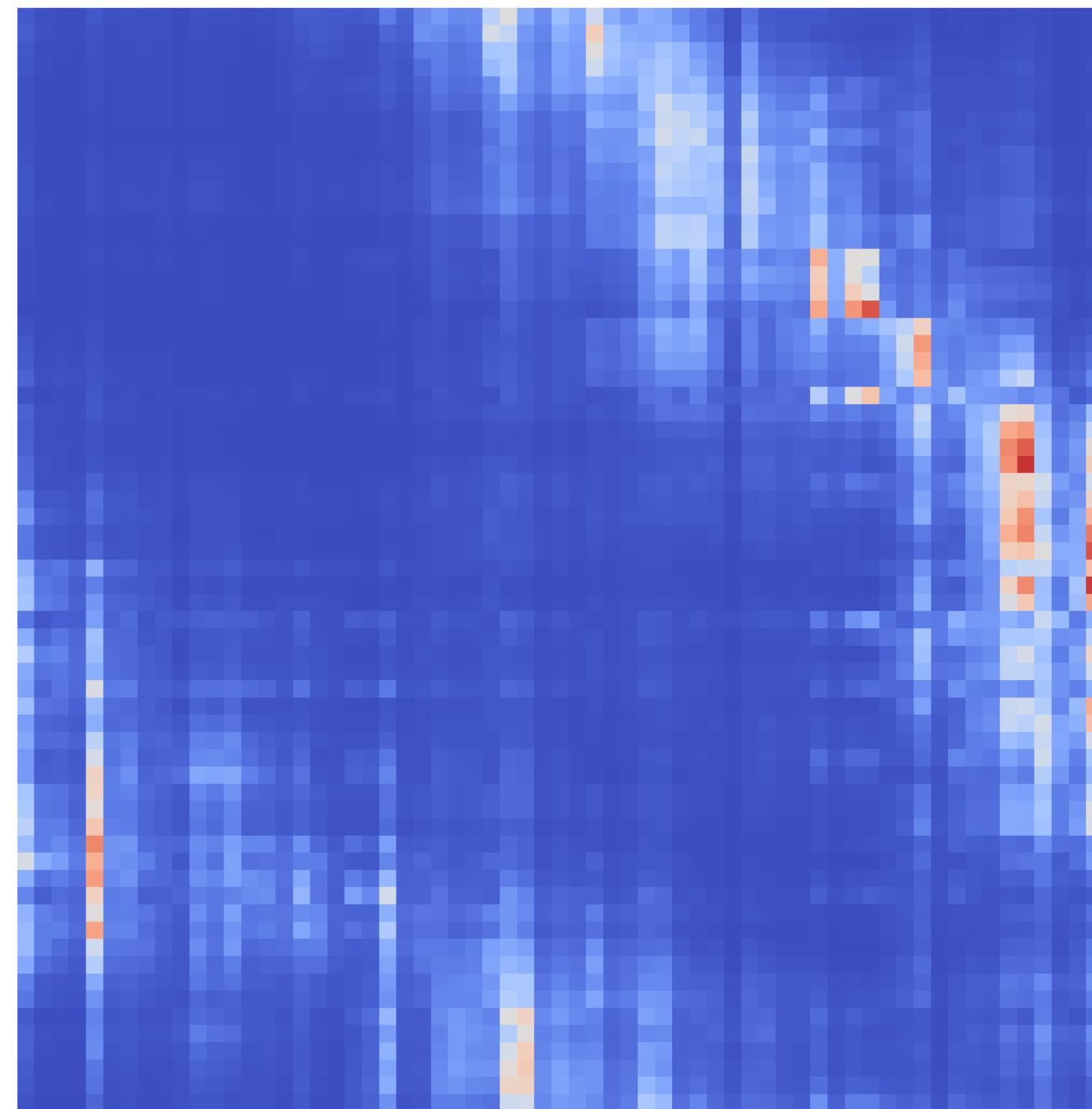
²University of Surrey

³Huawei Noah's Ark Lab

Multi-Resolution Approximation

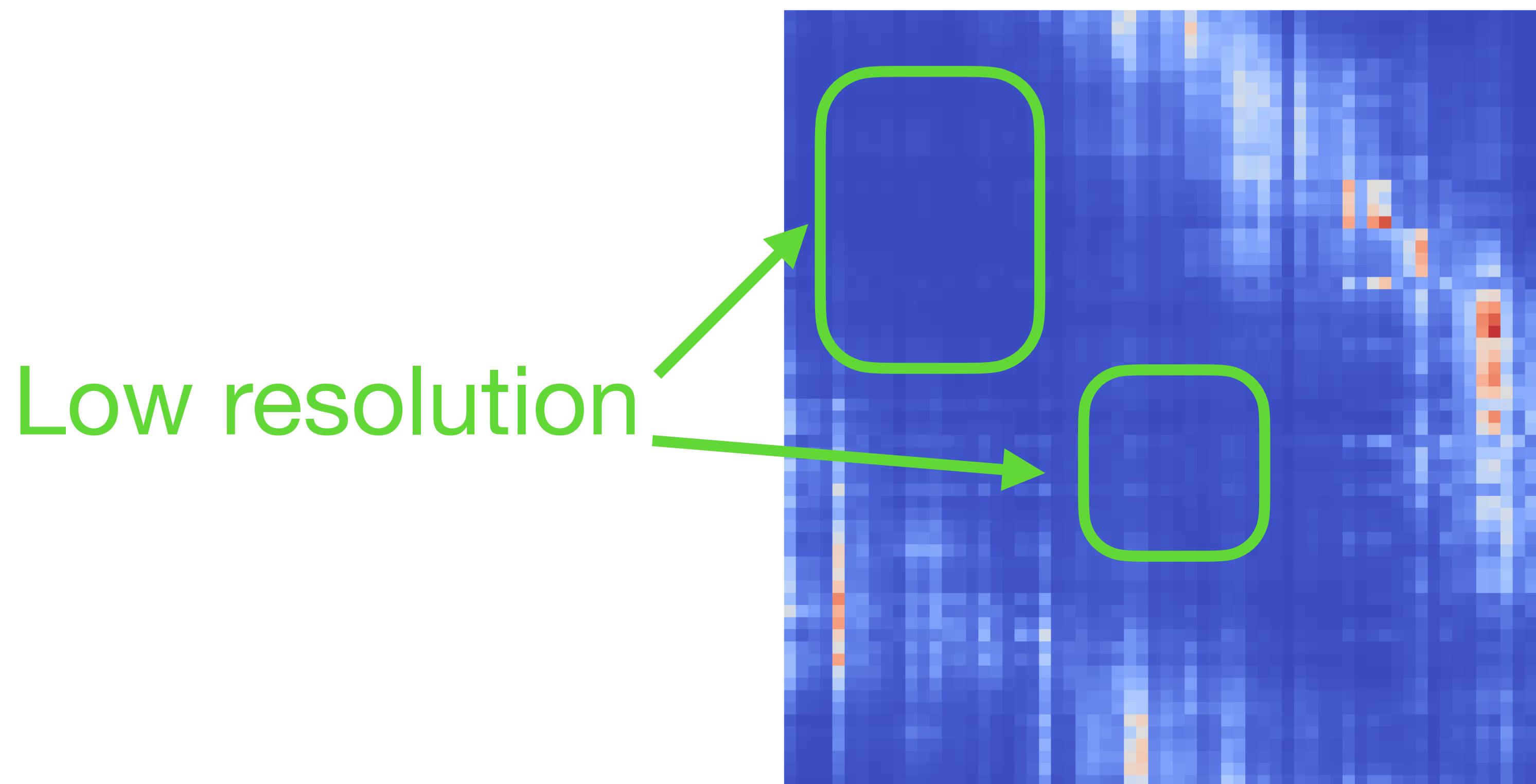
Multi-Resolution Approximation

Why?



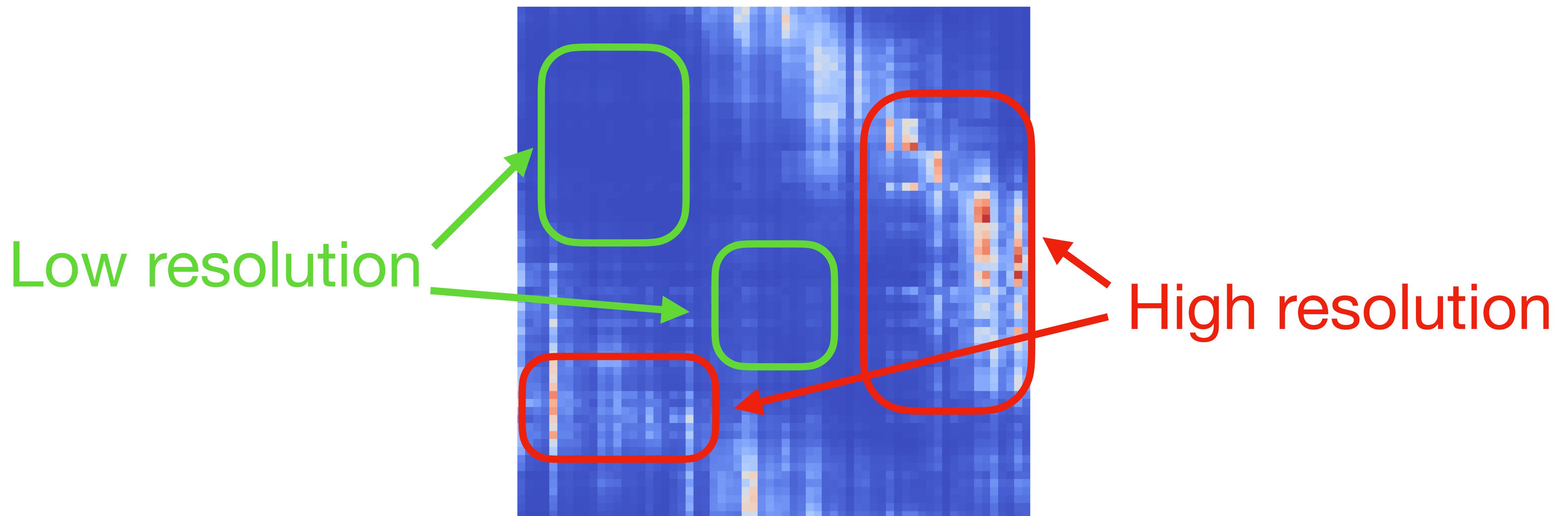
Multi-Resolution Approximation

Why?

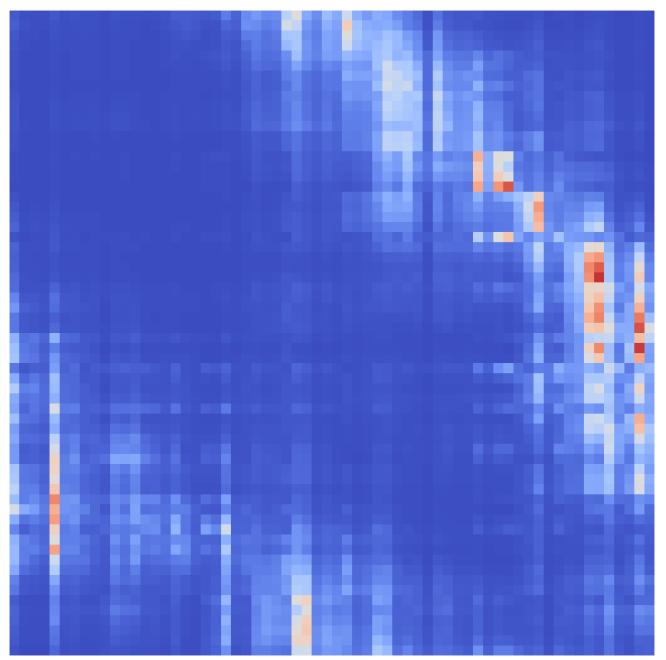


Multi-Resolution Approximation

Why?



Multi-Resolution Approximation Wavelets



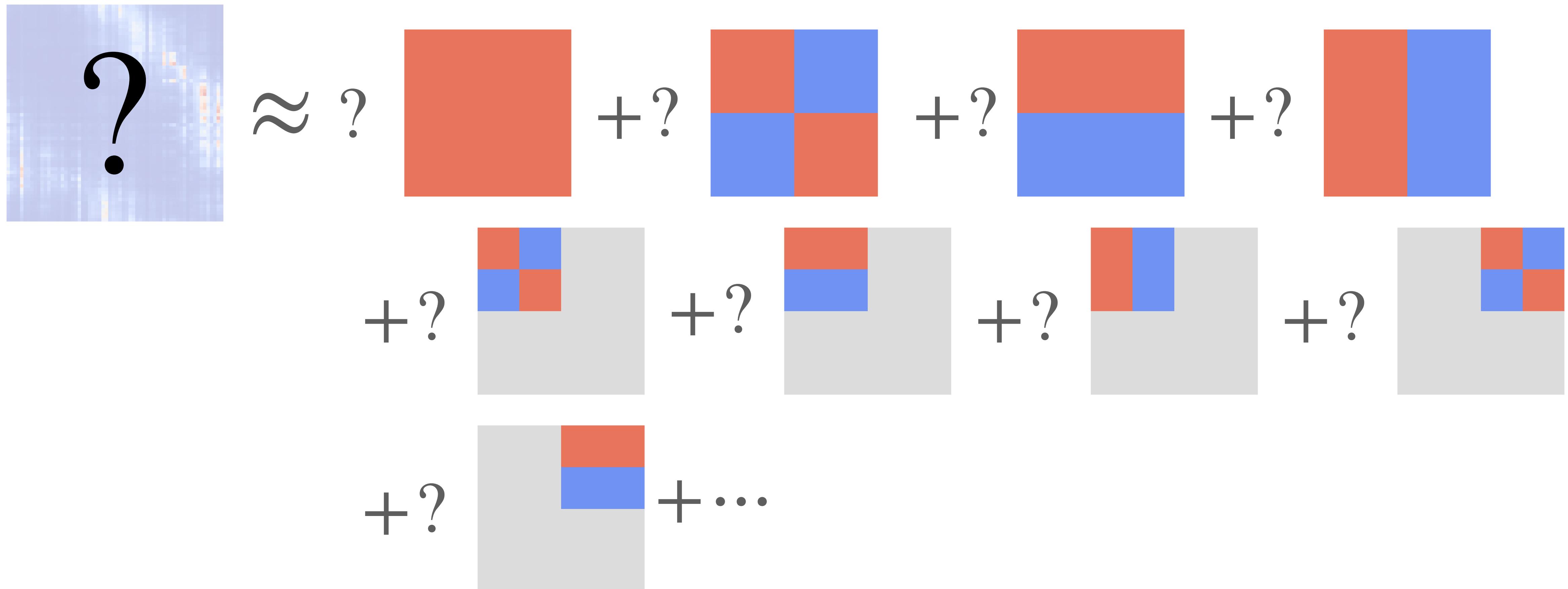
Multi-Resolution Approximation

Wavelets

$$\text{approx. image} \approx \alpha_1 \text{ (red block)} + \alpha_2 \text{ (blue and red blocks)} + \alpha_3 \text{ (blue and red blocks)} + \alpha_4 \text{ (red and blue blocks)}$$
$$+ \alpha_5 \text{ (grey block with red and blue corners)} + \alpha_6 \text{ (grey block with red and blue corners)} + \alpha_7 \text{ (grey block with red and blue corners)} + \alpha_8 \text{ (grey block with red and blue corners)}$$
$$+ \alpha_9 \text{ (grey block with red and blue corners)} + \dots$$

Multi-Resolution Approximation

Wavelets



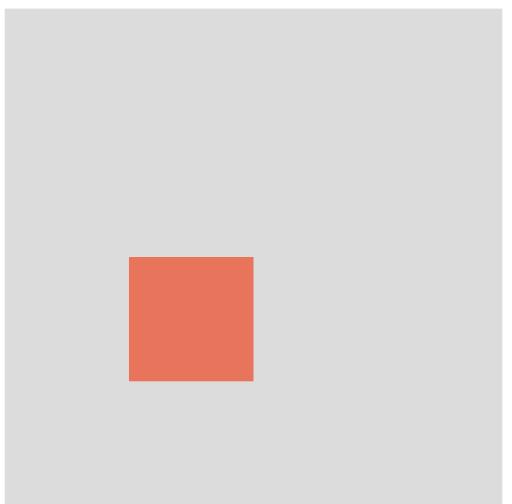
Multi-Resolution Approximation

Wavelets

$$\text{Image} \approx \mu_1 \begin{array}{c} \text{Red} \\ \text{Grey} \end{array} + \mu_2 \begin{array}{c} \text{Grey} \\ \text{Red} \end{array} + \mu_3 \begin{array}{c} \text{Grey} \\ \text{Red} \\ \text{Grey} \end{array} + \mu_4 \begin{array}{c} \text{Grey} \\ \text{Grey} \\ \text{Red} \end{array} + \mu_5 \begin{array}{c} \text{Grey} \\ \text{Red} \\ \text{Grey} \end{array} + \mu_6 \begin{array}{c} \text{Grey} \\ \text{Red} \\ \text{Grey} \\ \text{Grey} \end{array} + \mu_7 \begin{array}{c} \text{Grey} \\ \text{Red} \\ \text{Grey} \\ \text{Grey} \\ \text{Grey} \end{array} + \mu_8 \begin{array}{c} \text{Grey} \\ \text{Grey} \\ \text{Red} \\ \text{Grey} \\ \text{Grey} \end{array} + \mu_9 \begin{array}{c} \text{Red} \\ \text{Grey} \\ \text{Grey} \\ \text{Grey} \\ \text{Grey} \\ \text{Grey} \end{array} + \dots$$

Multi-Resolution Approximation

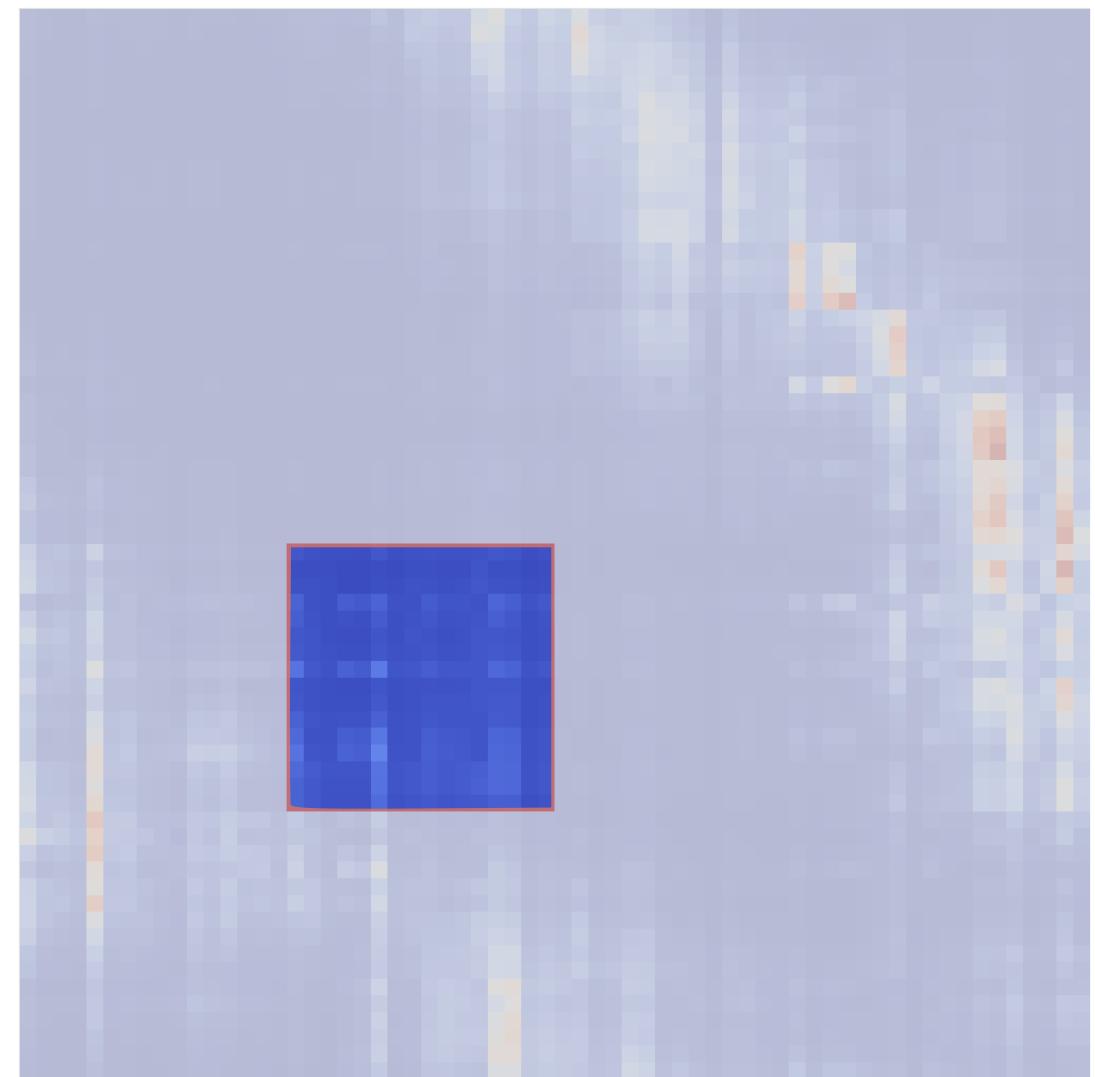
Coefficient Approximation

$$\dots + \mu_5$$

$$+ \dots$$

Multi-Resolution Approximation

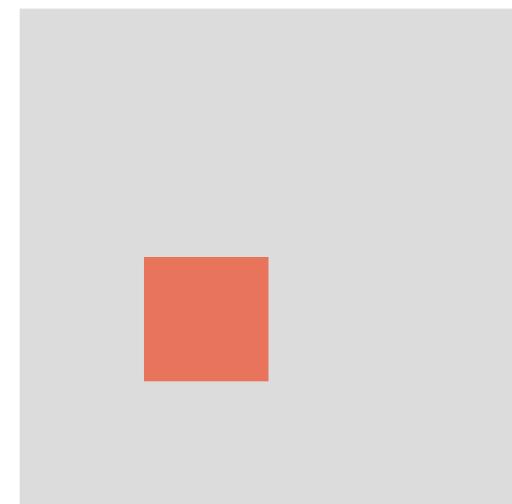
Coefficient Approximation

$$\mu_5 = Avg($$



)

$$\dots + \mu_5$$

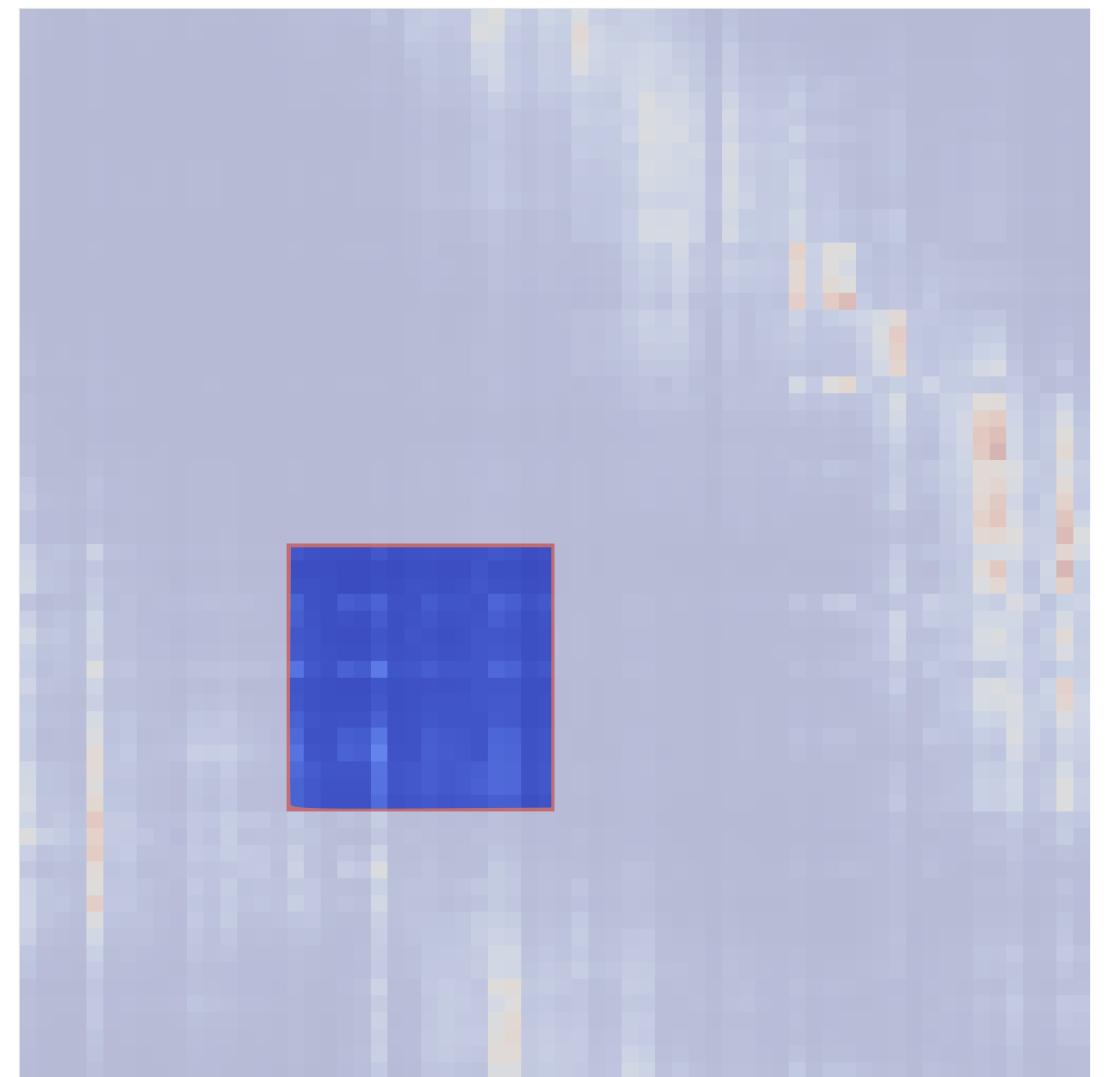


$$+ \dots$$

Multi-Resolution Approximation

Coefficient Approximation

$$\mu_5 = Avg($$



)

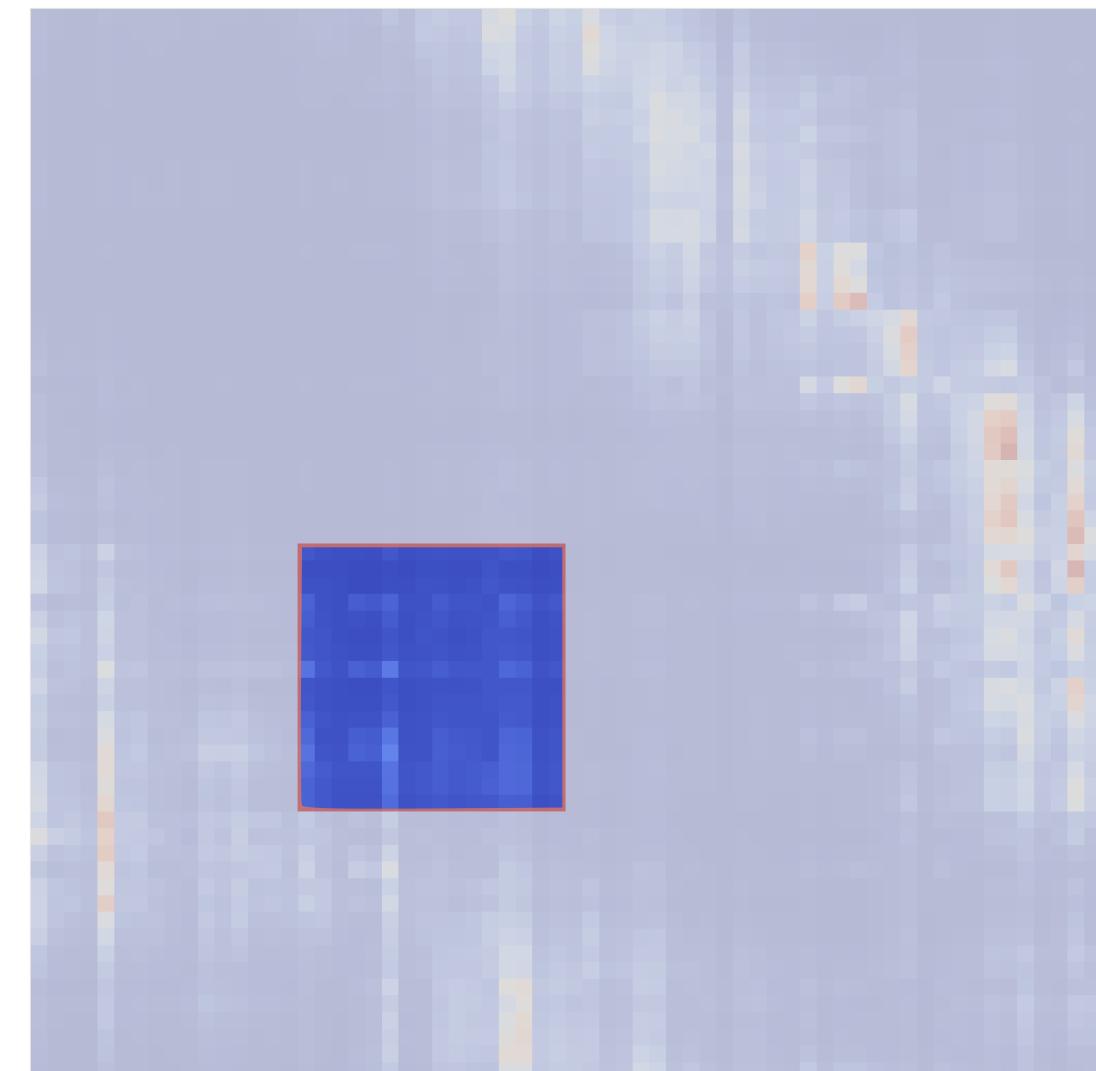
$$\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} \exp(QK^T)_{i,j}$$

$$\dots + \mu_5 \quad + \dots$$

Multi-Resolution Approximation

Coefficient Approximation

$$\mu_5 = Avg($$



)

$$\dots + \mu_5 \quad + \dots$$

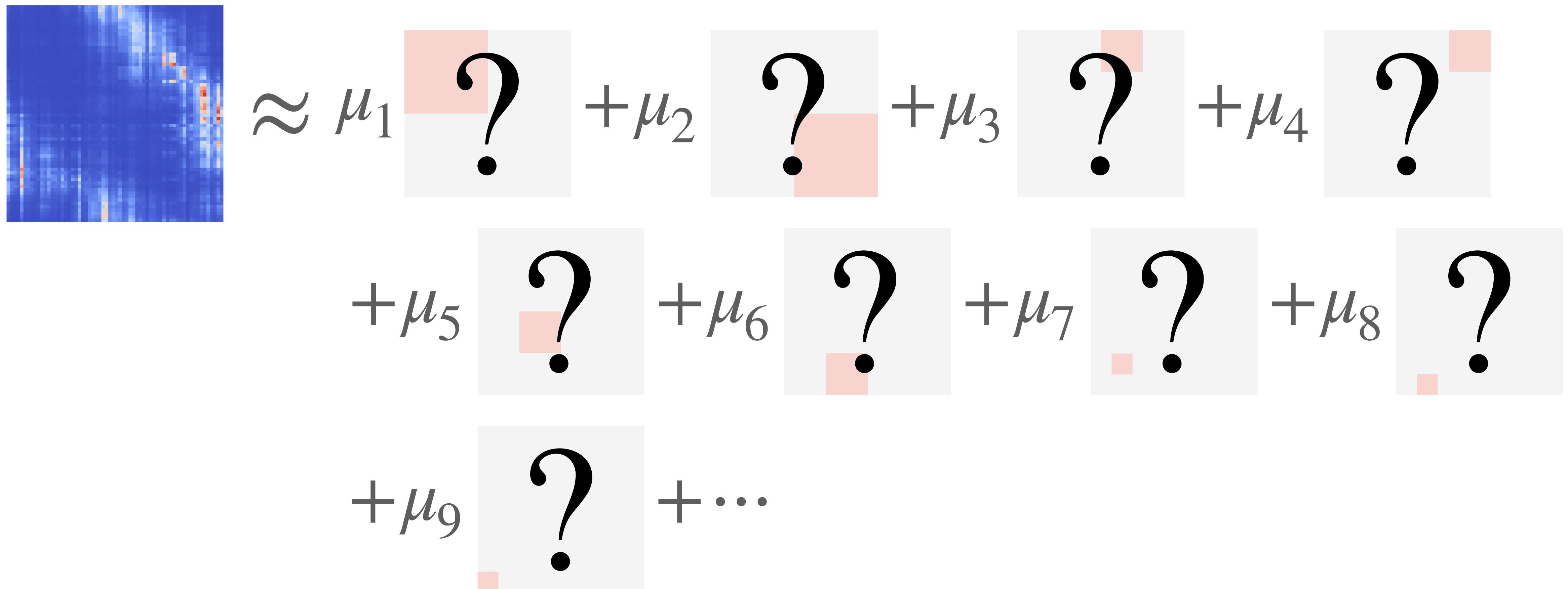
$$\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} \exp(QK^T)_{i,j}$$

Efficient

$$\exp\left(\frac{1}{N} \sum_{(B_5)_{i,j} \neq 0} (QK^T)_{i,j}\right)$$

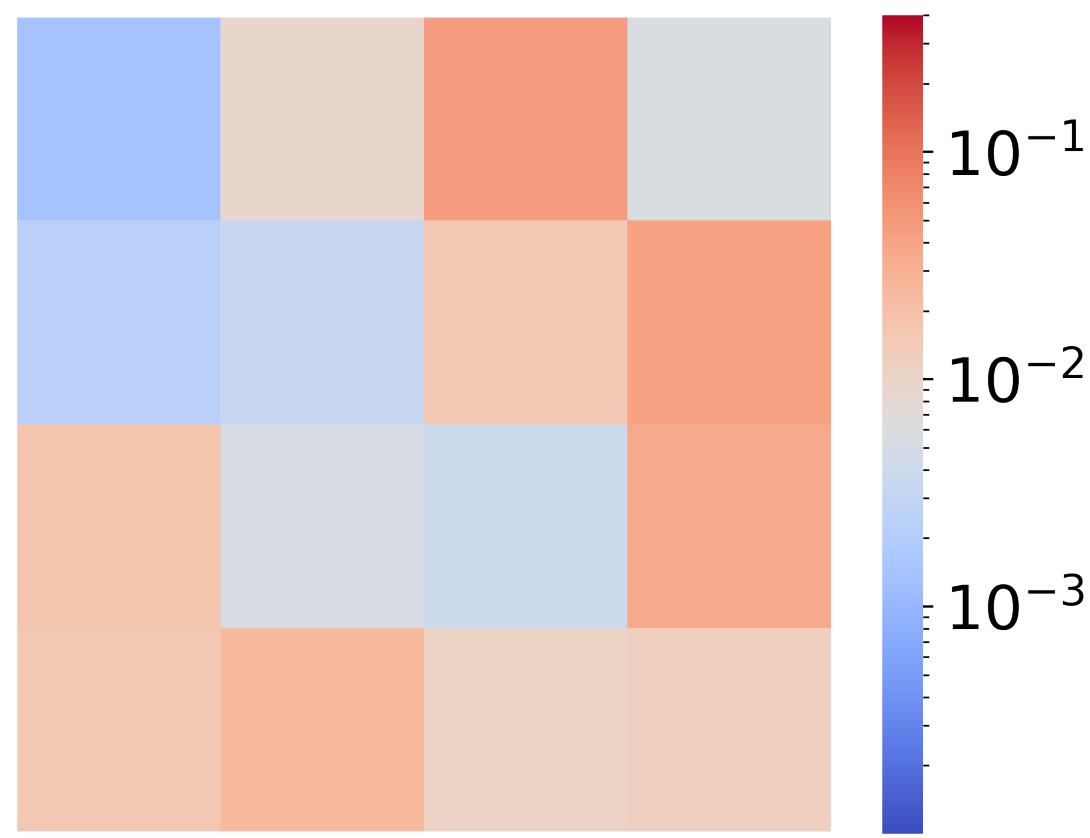
Multi-Resolution Approximation

Component Selection


$$\approx \mu_1 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_2 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_3 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_4 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_5 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_6 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_7 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_8 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \mu_9 \begin{matrix} \text{?} \\ \text{?} \end{matrix} + \dots$$

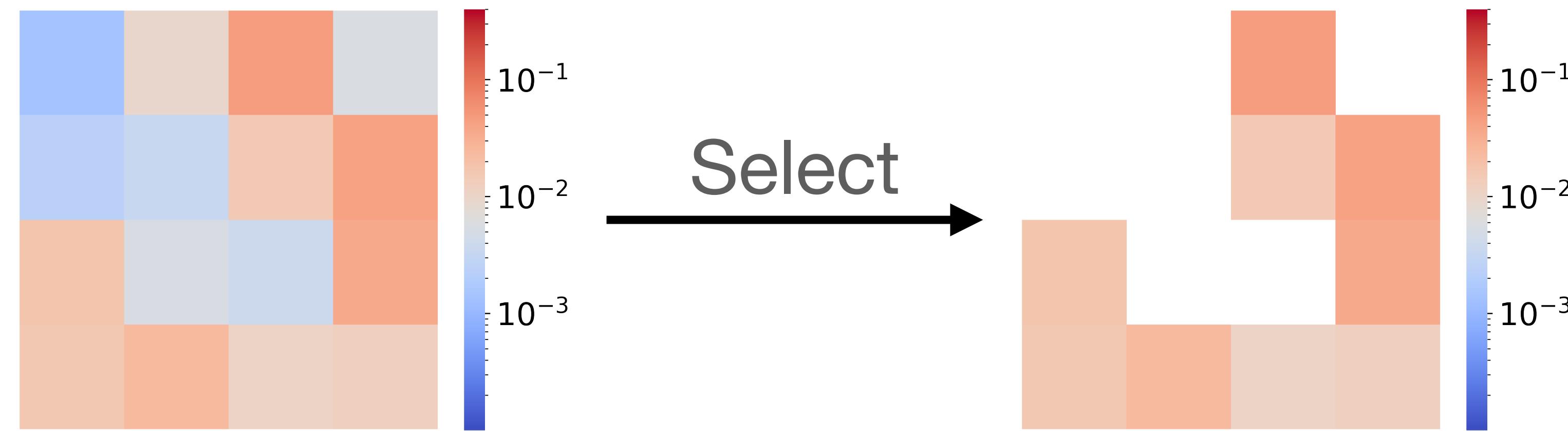
Multi-Resolution Approximation

Illustration

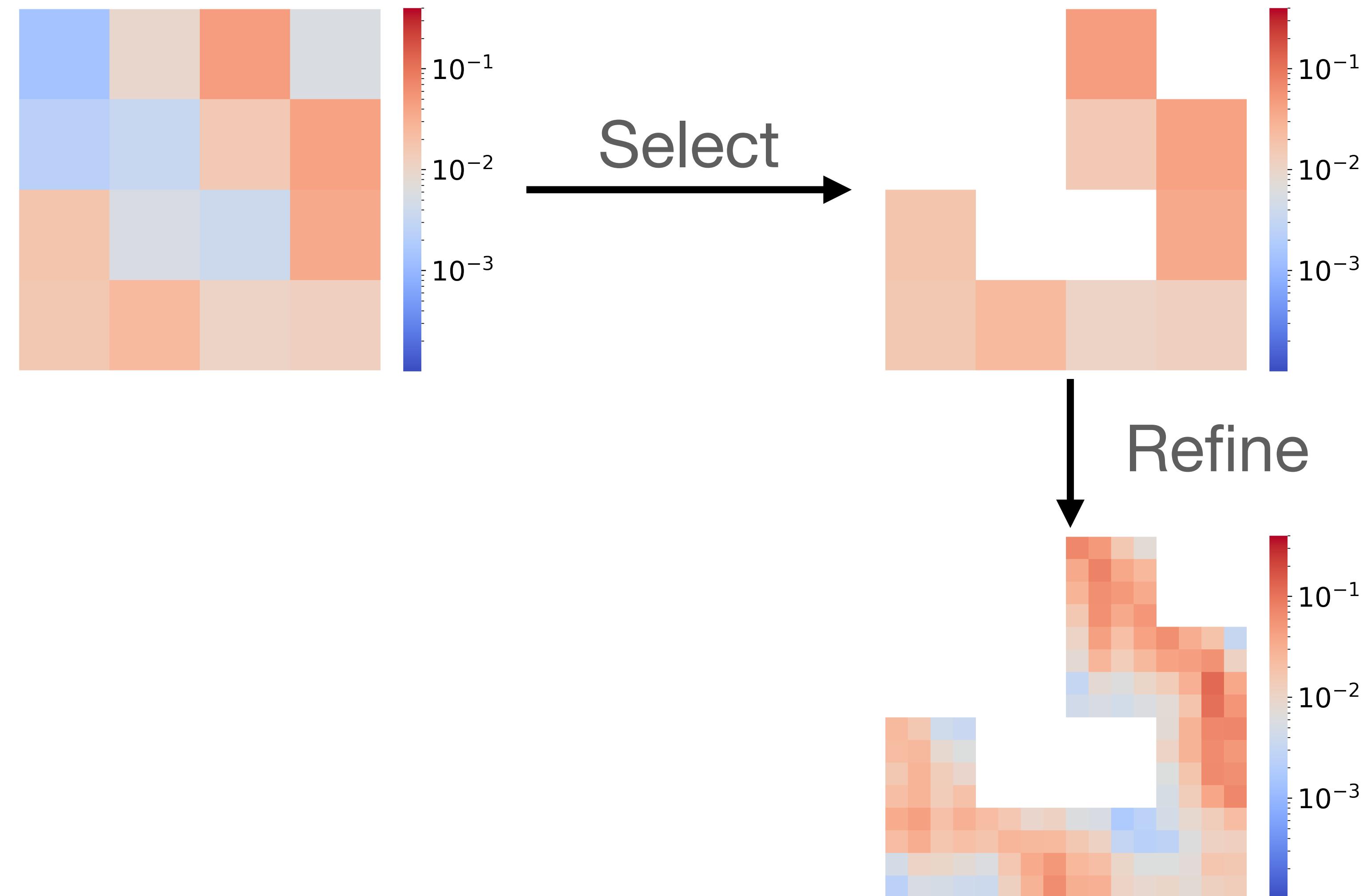


Multi-Resolution Approximation

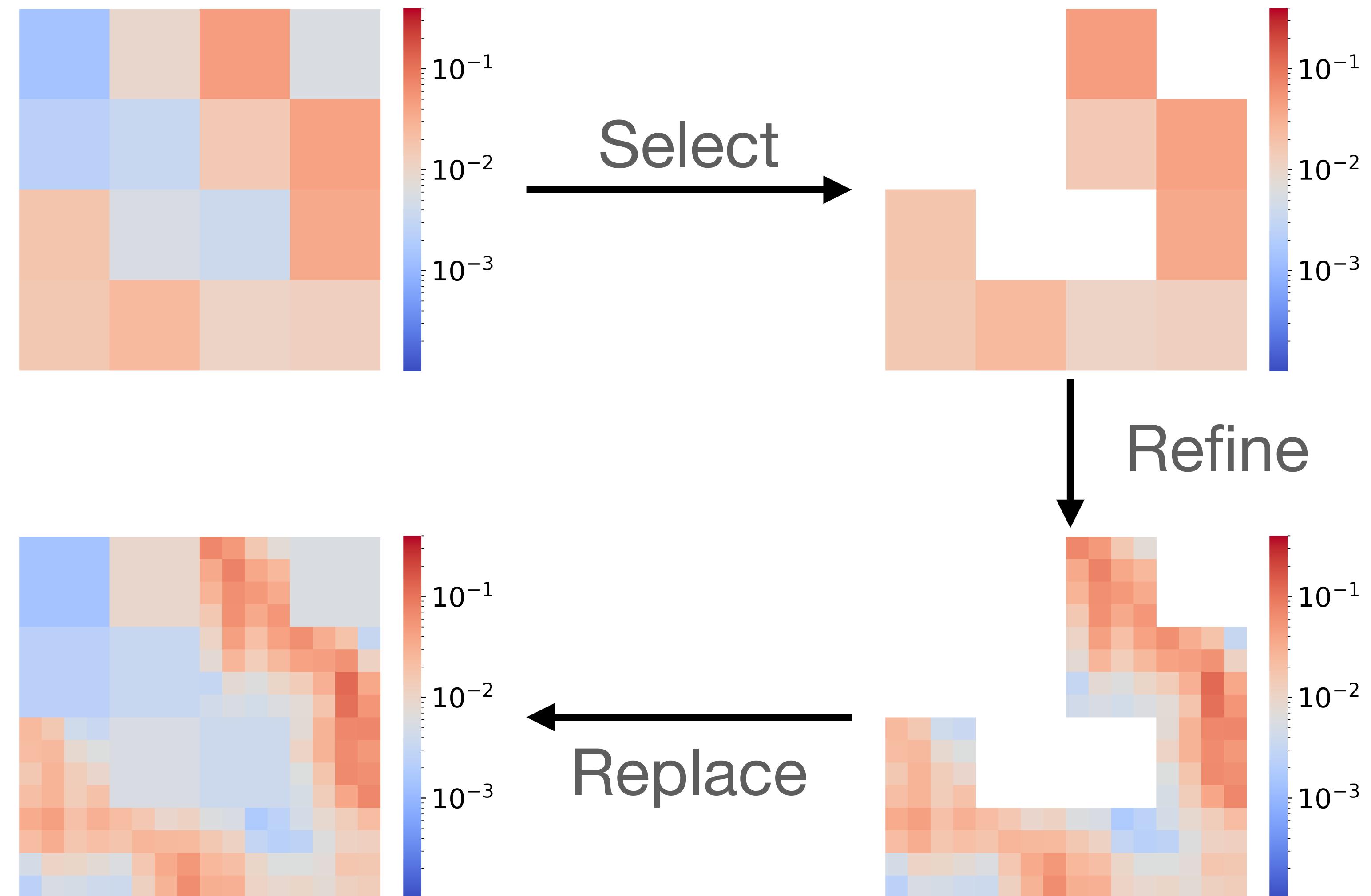
Illustration



Multi-Resolution Approximation Illustration

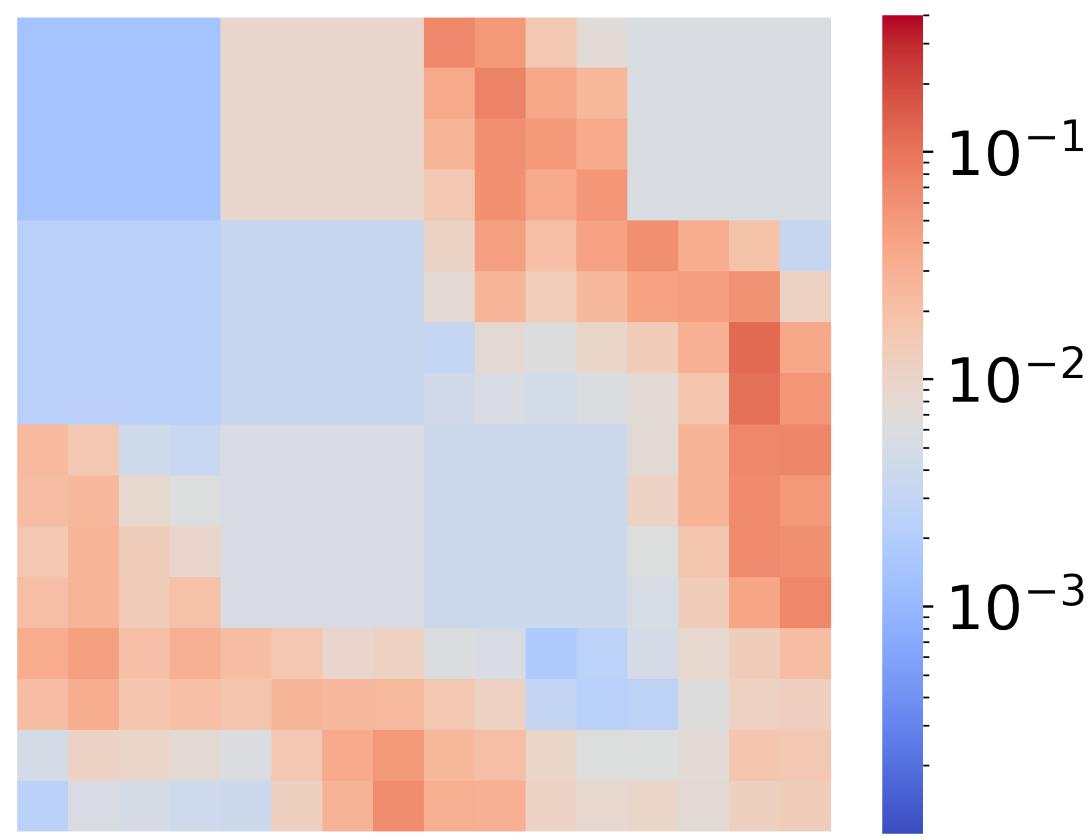


Multi-Resolution Approximation Illustration

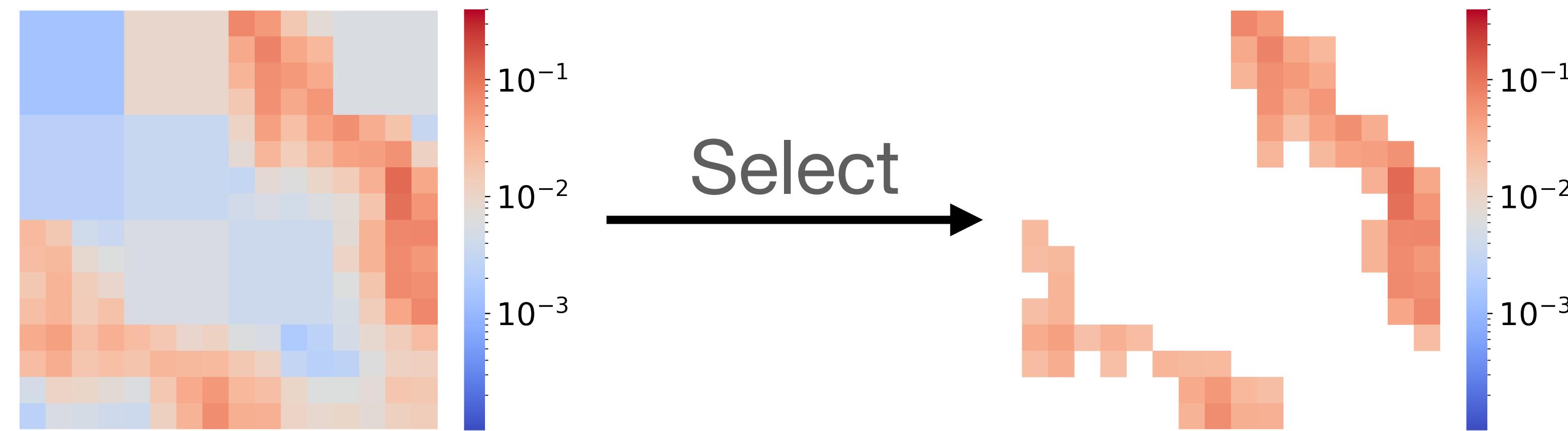


Multi-Resolution Approximation

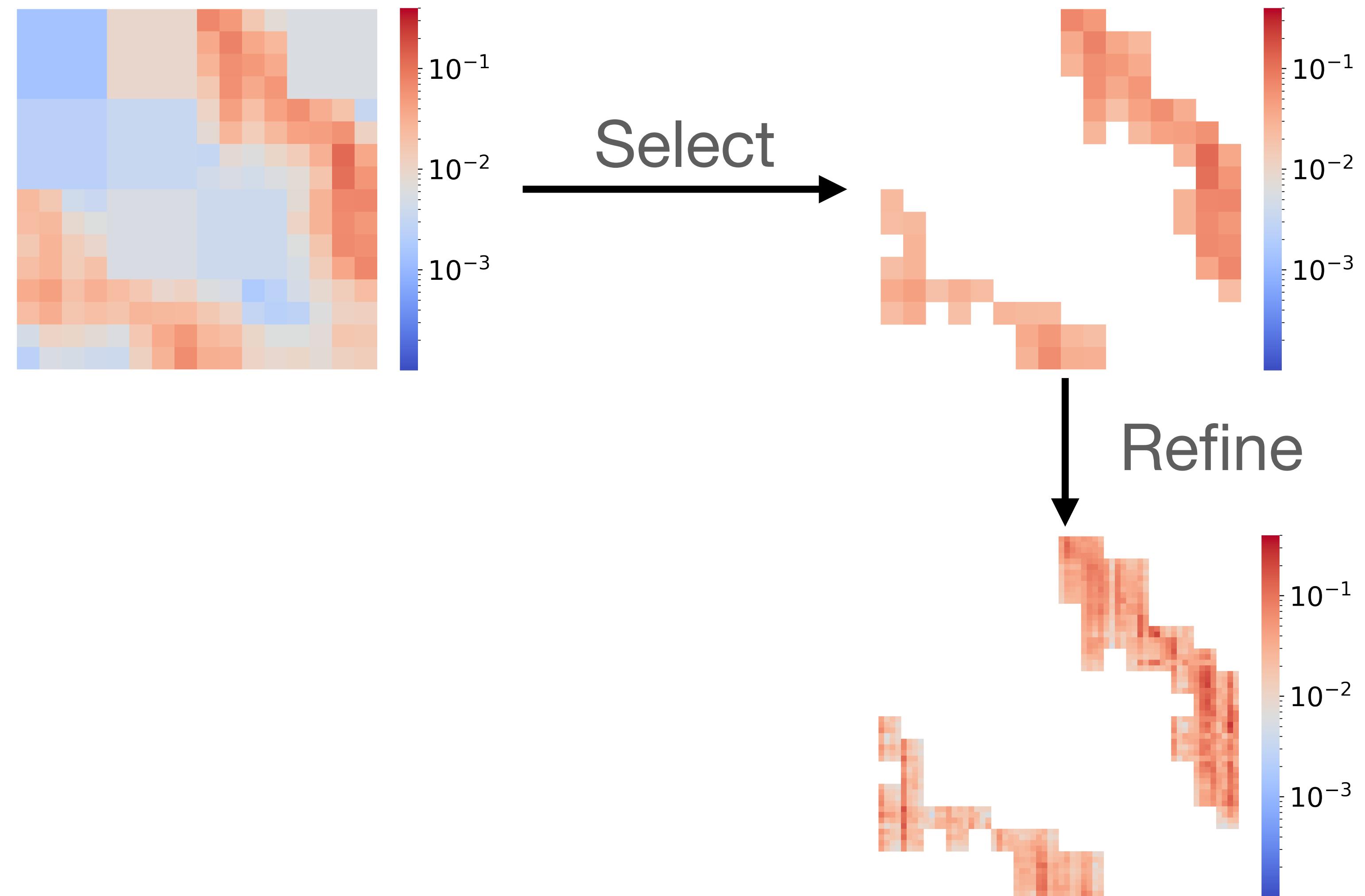
Illustration



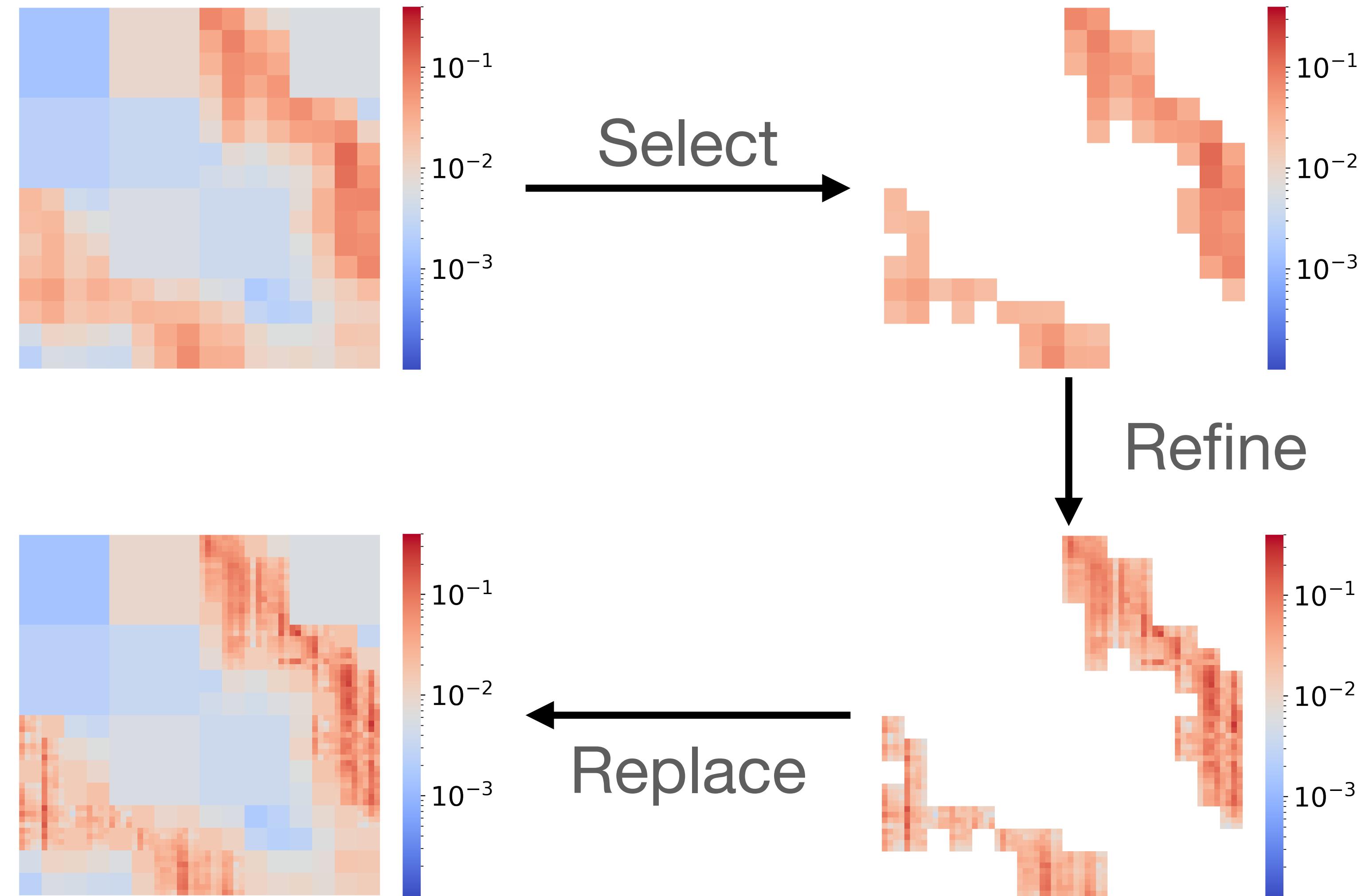
Multi-Resolution Approximation Illustration



Multi-Resolution Approximation Illustration

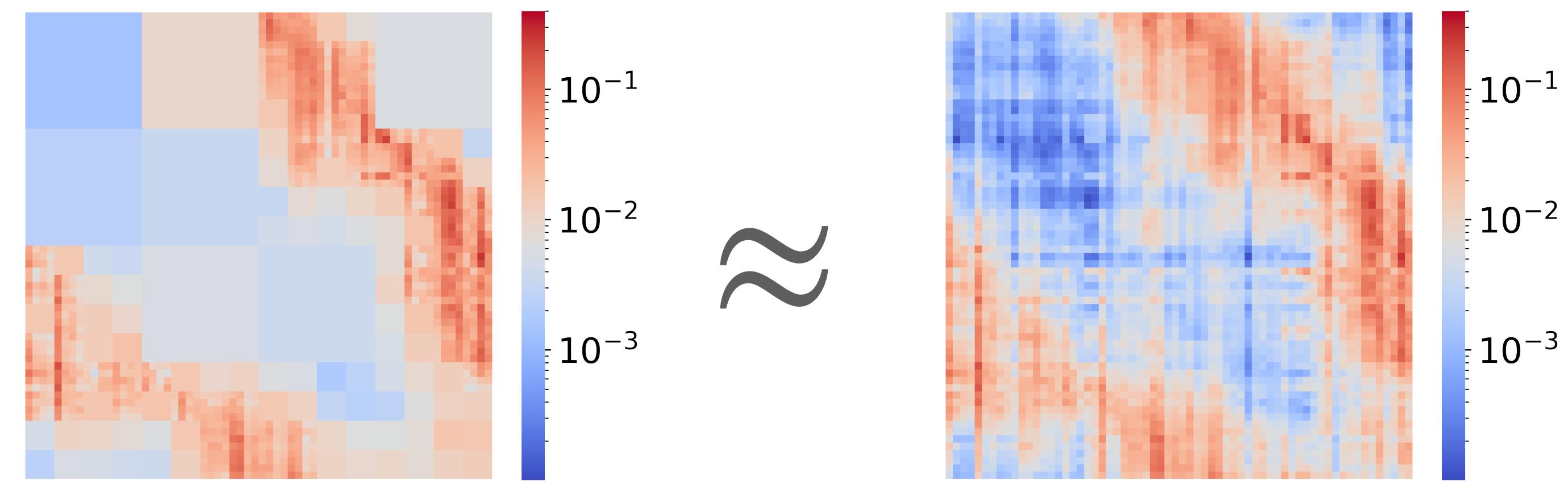


Multi-Resolution Approximation Illustration



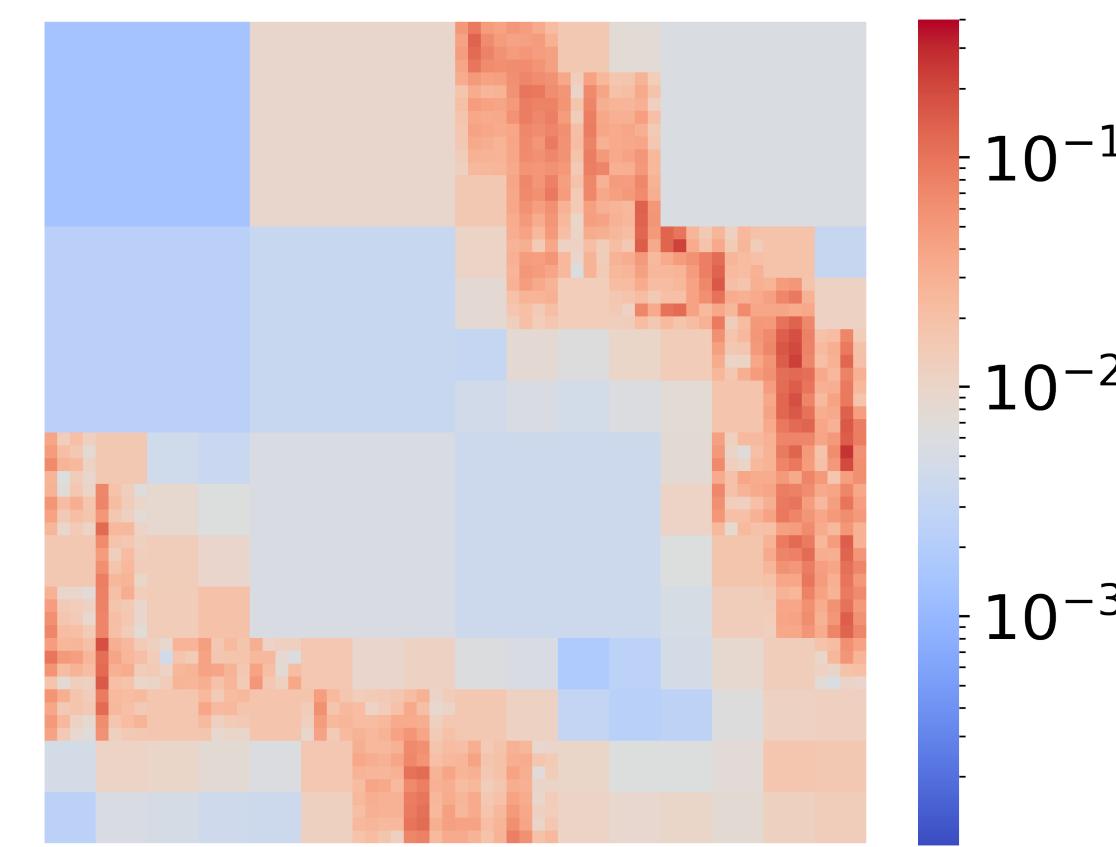
Multi-Resolution Approximation

Approximation Result

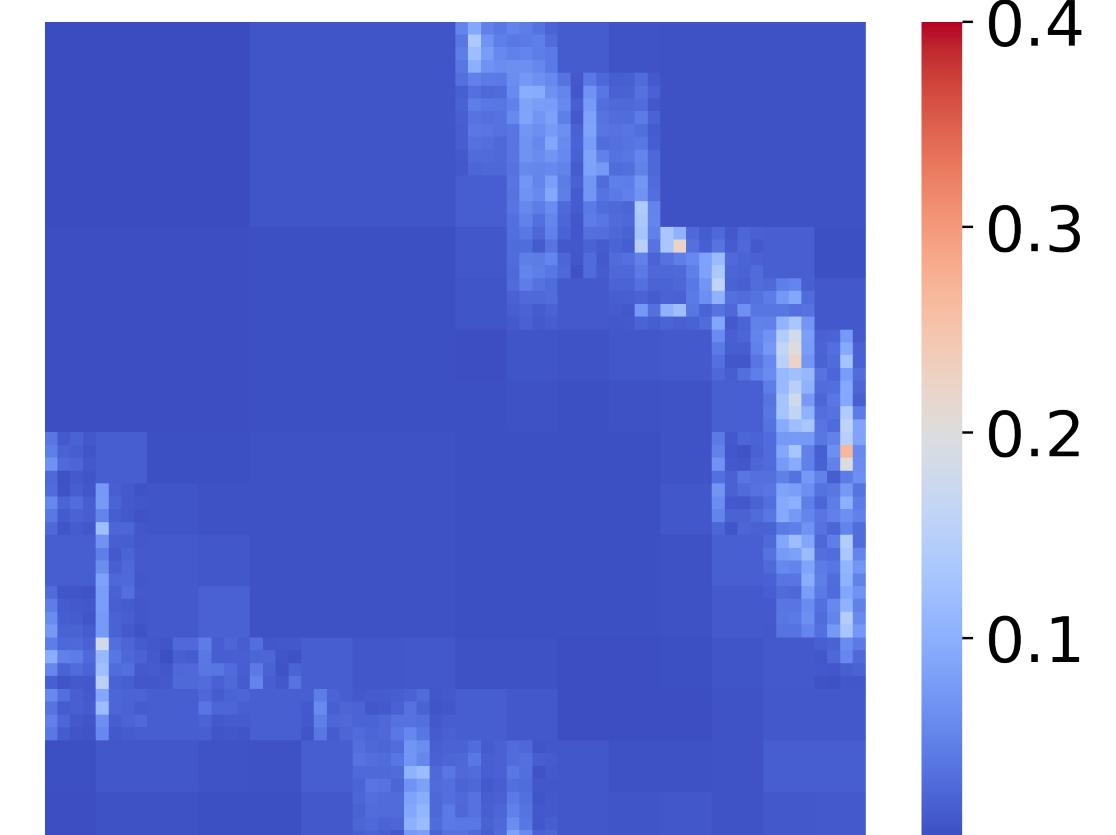
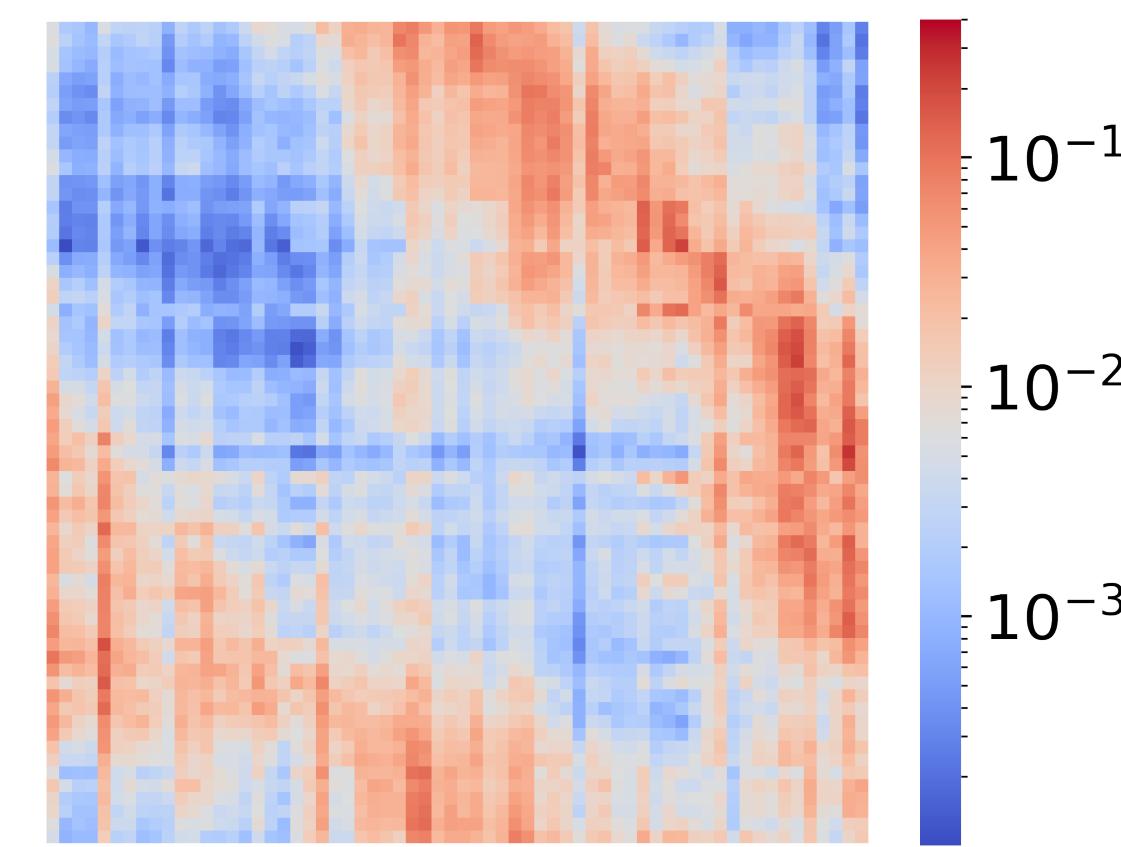


Multi-Resolution Approximation

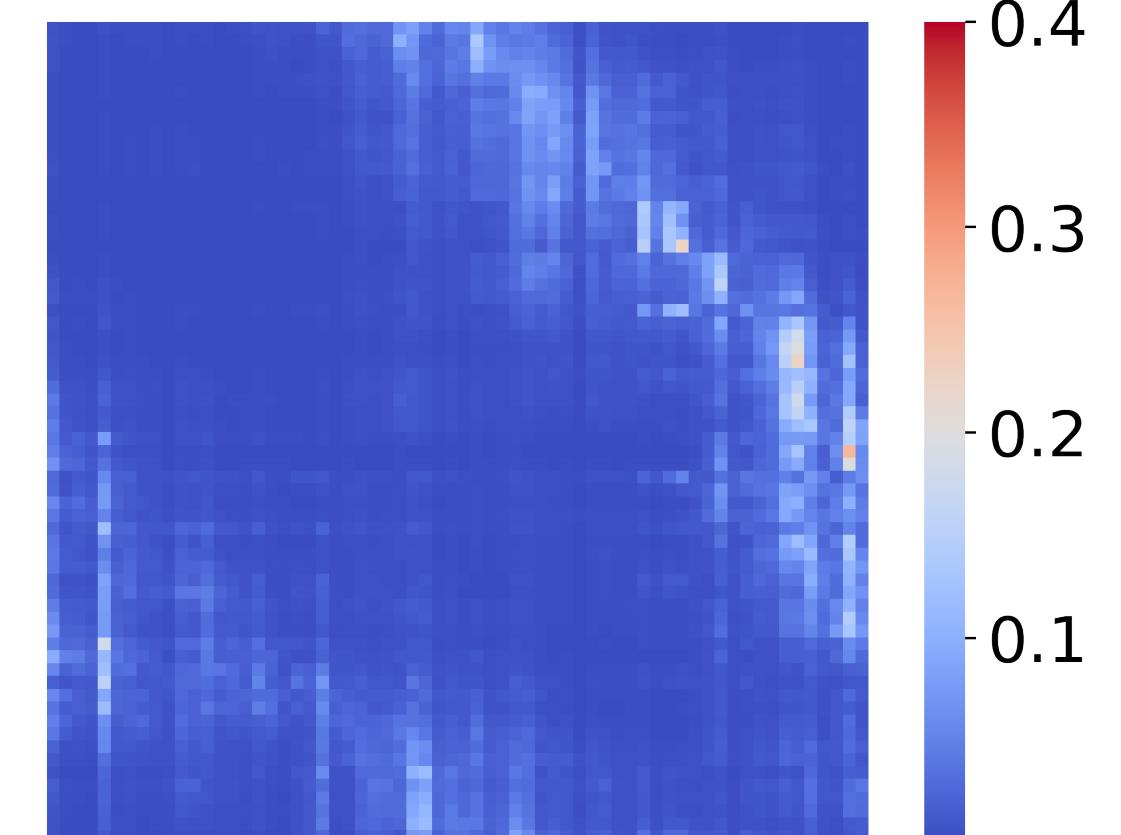
Approximation Result



\approx



\approx

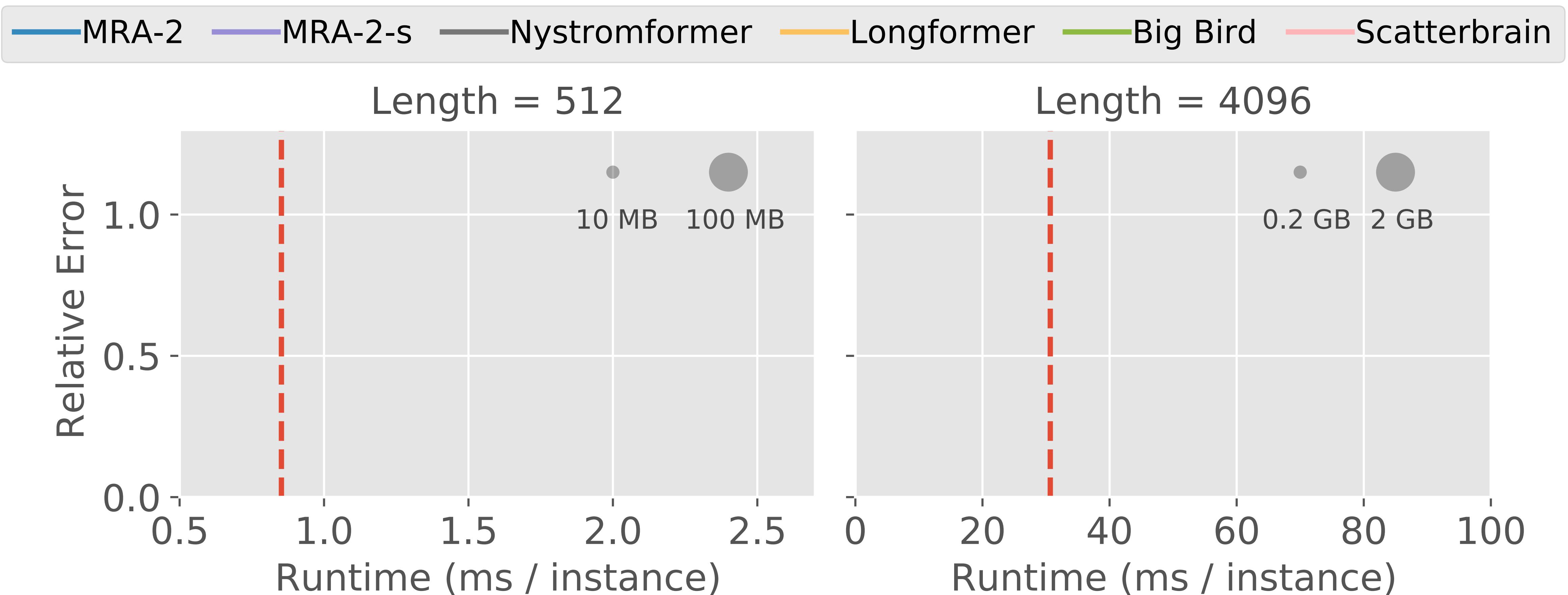


Evaluations

Efficiency versus Approximation

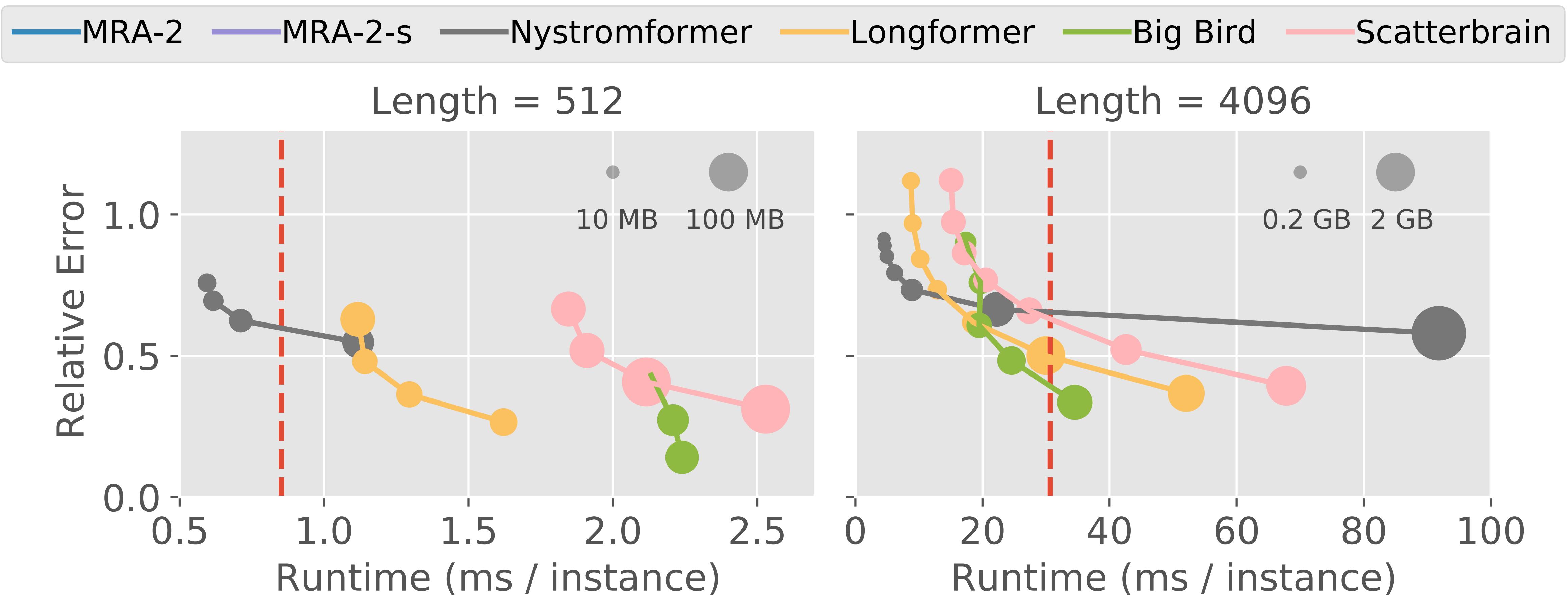
Evaluations

Efficiency v.s. Approximation



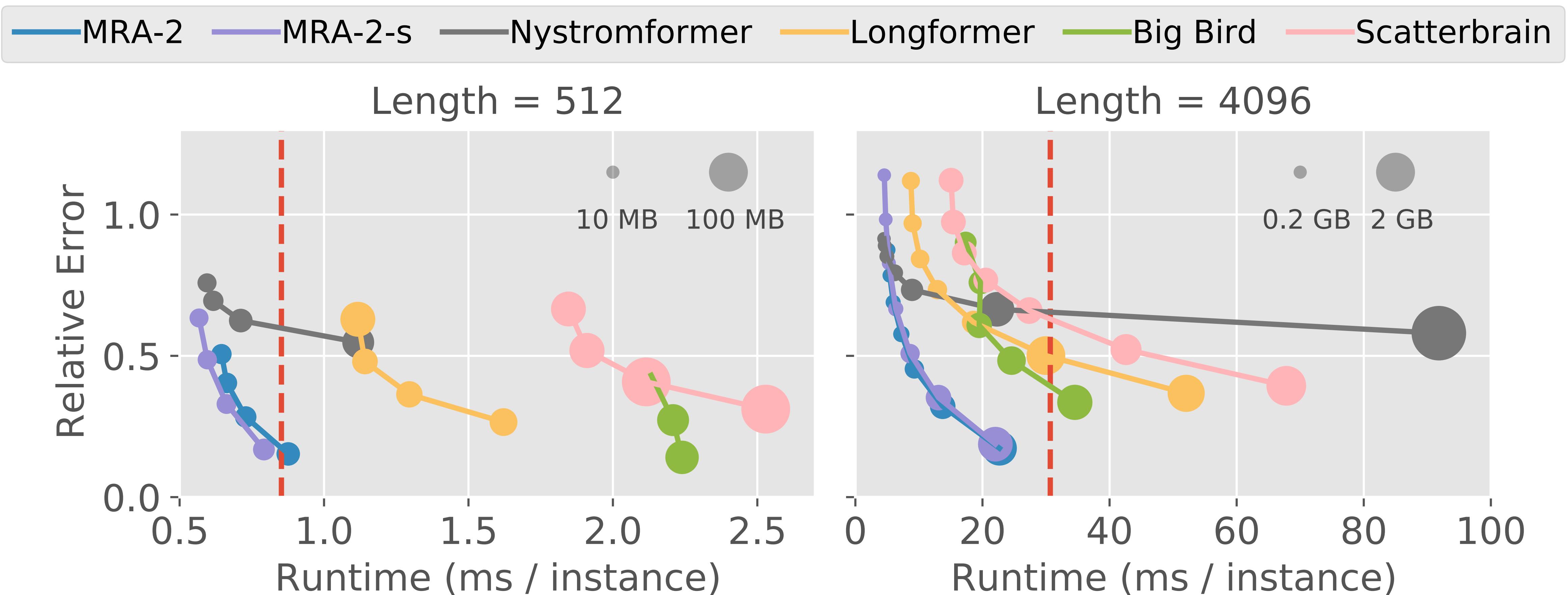
Evaluations

Efficiency v.s. Approximation



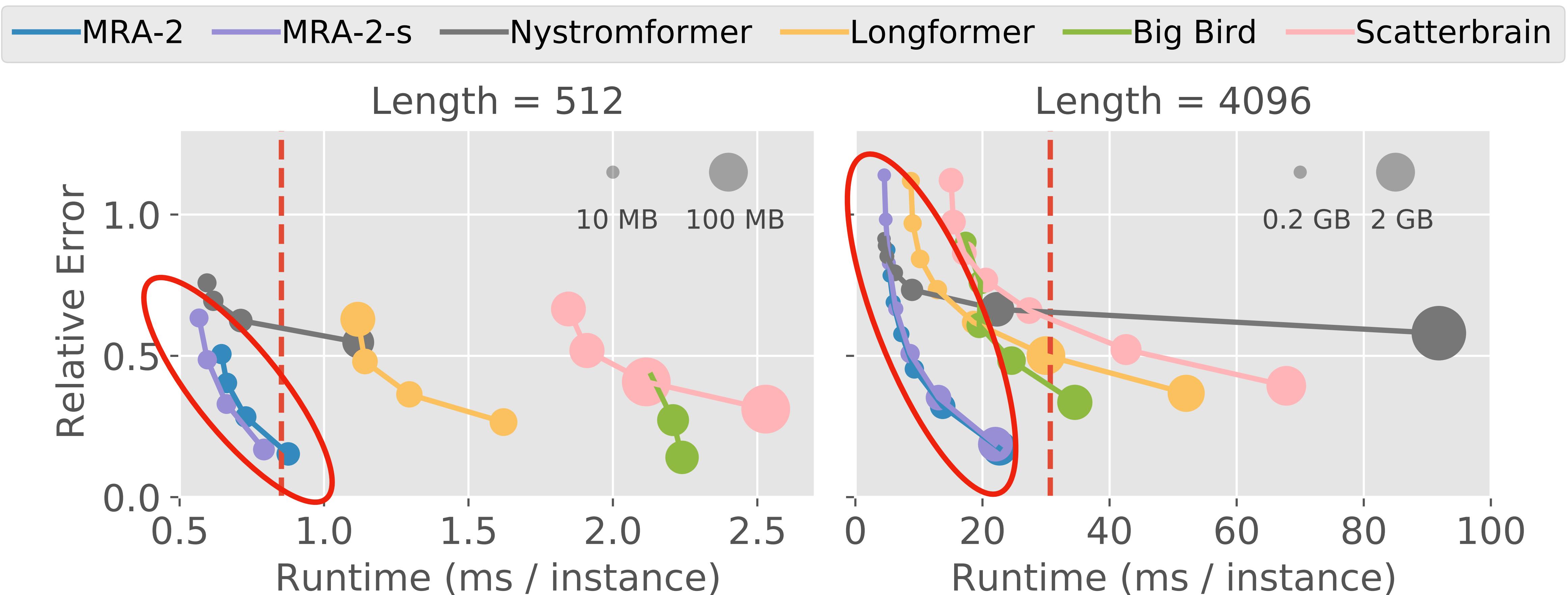
Evaluations

Efficiency v.s. Approximation



Evaluations

Efficiency v.s. Approximation

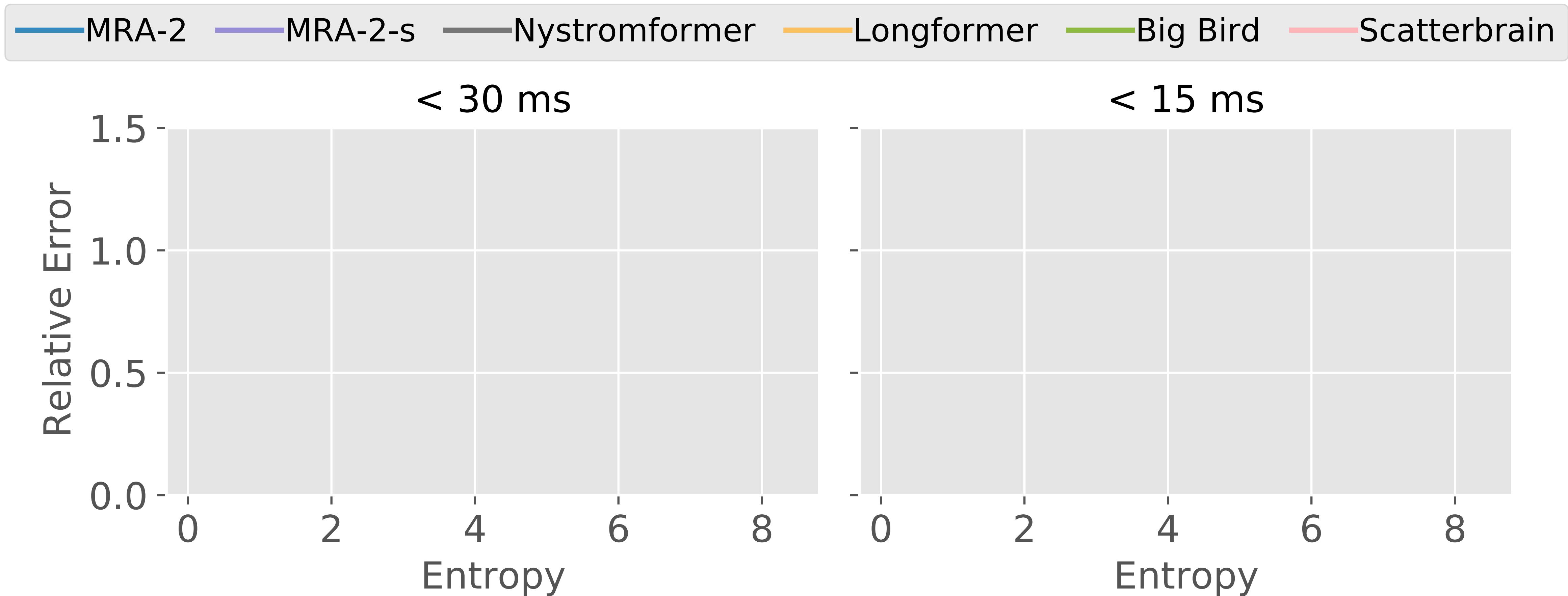


Evaluations

Entropy versus Approximation

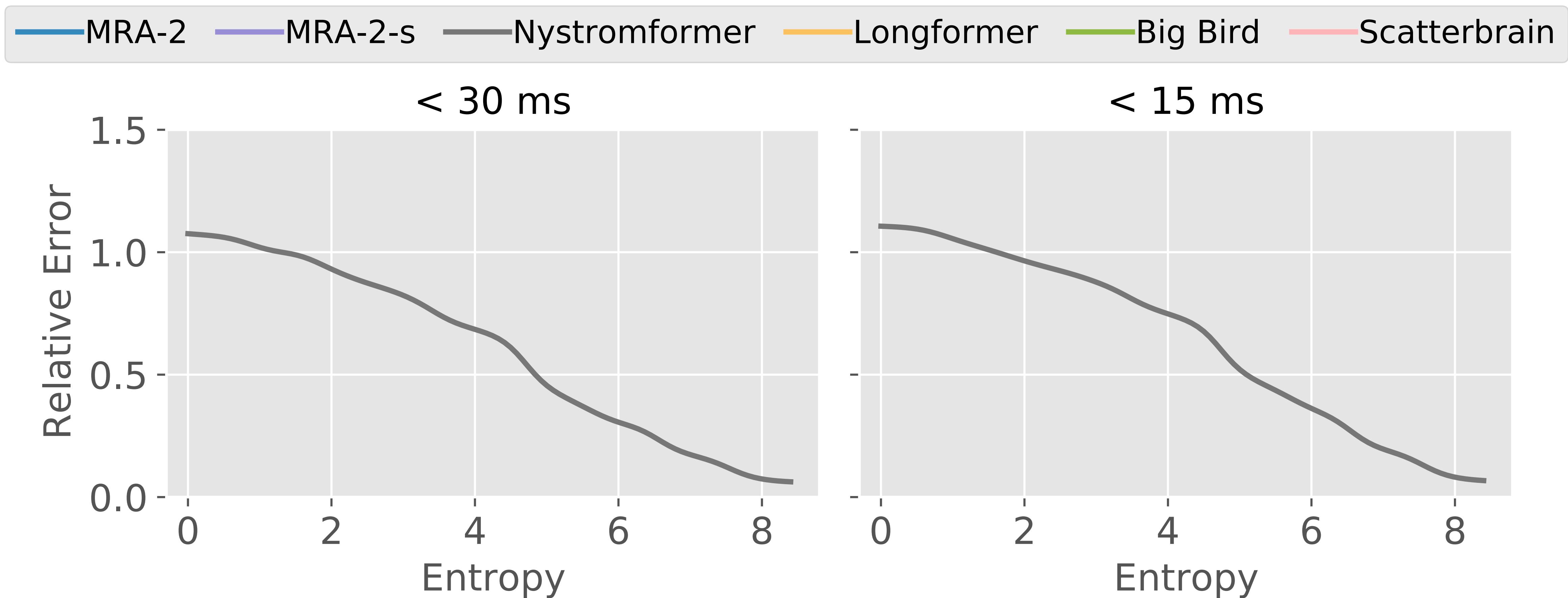
Evaluations

Entropy v.s. Approximation



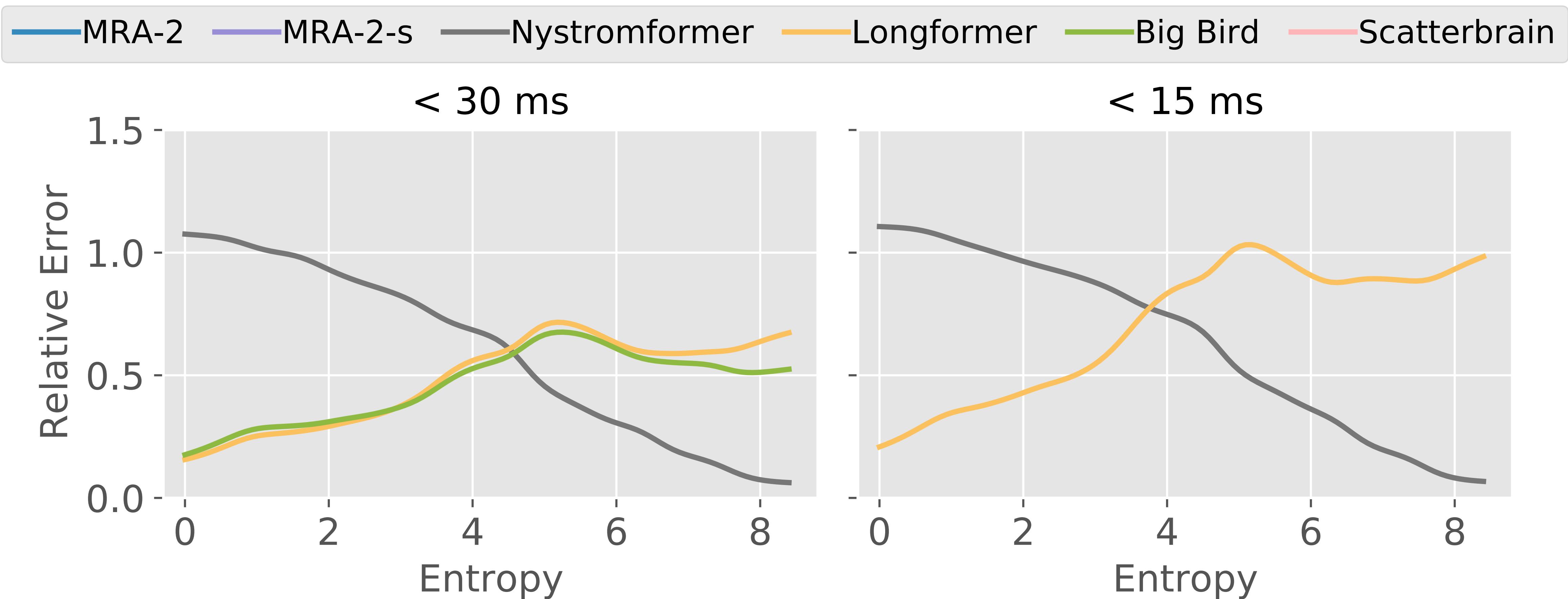
Evaluations

Entropy v.s. Approximation



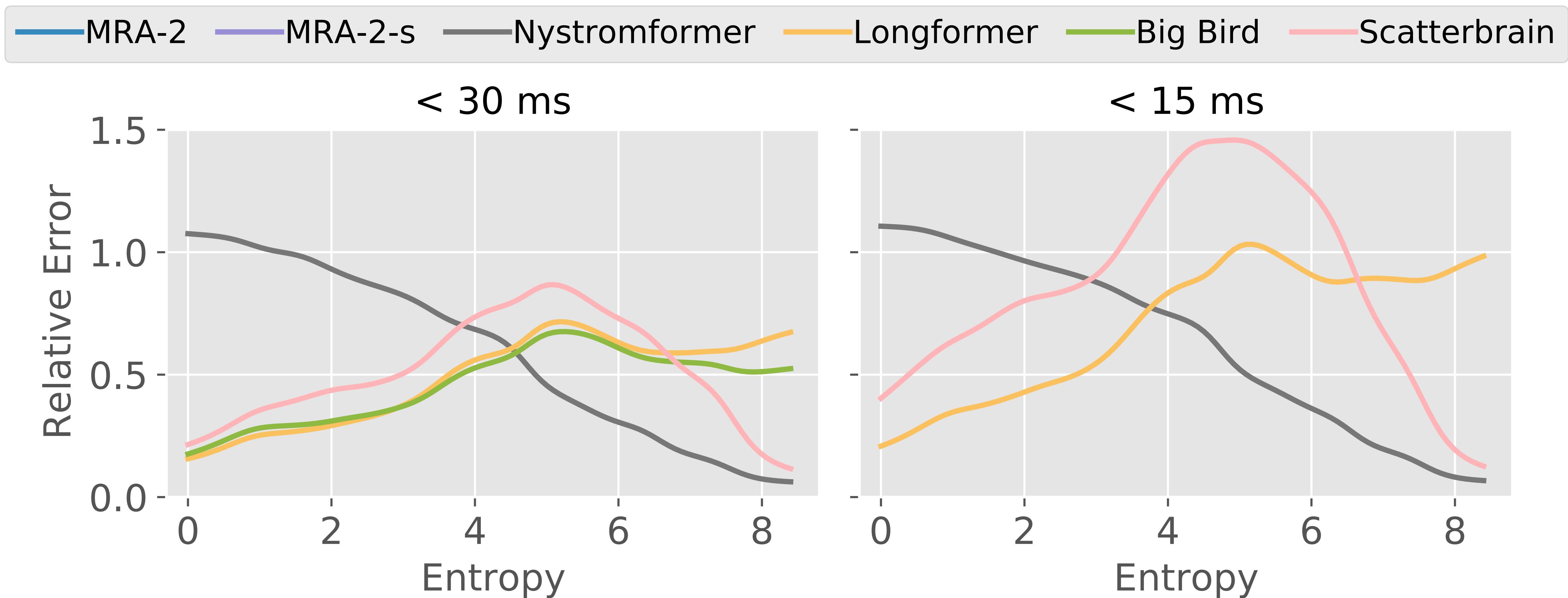
Evaluations

Entropy v.s. Approximation



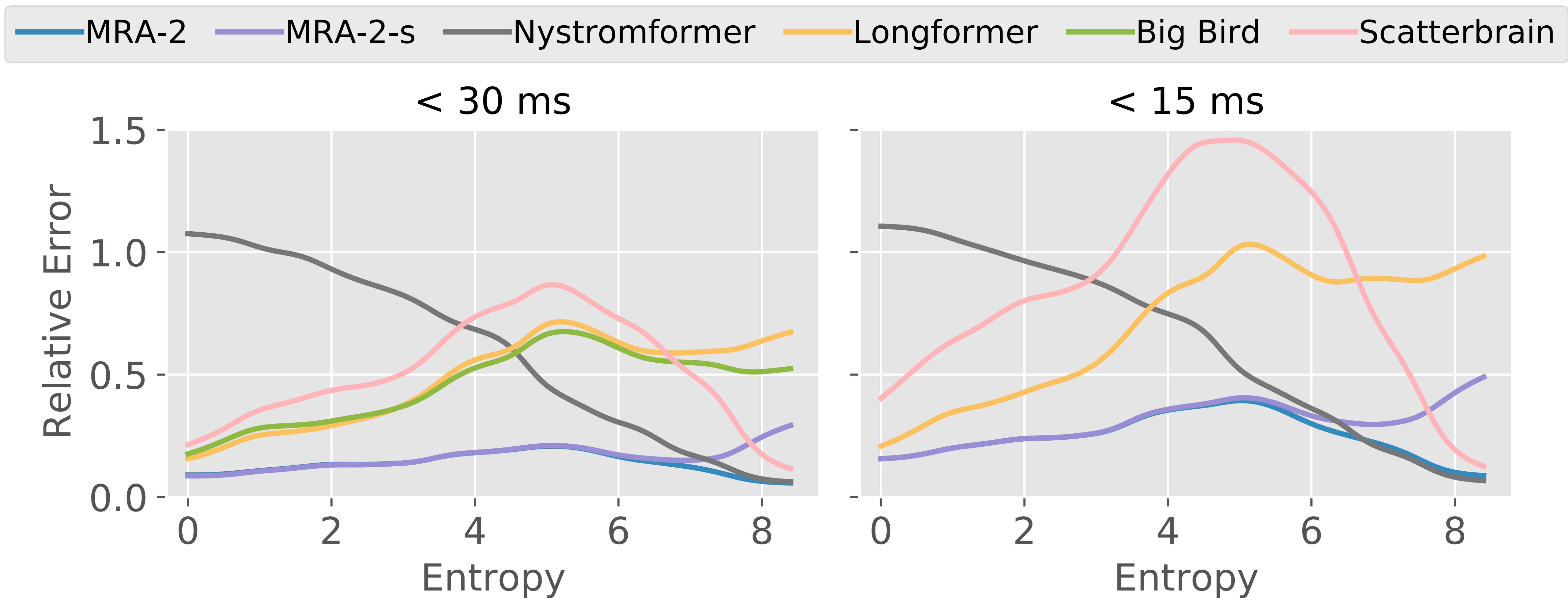
Evaluations

Entropy v.s. Approximation



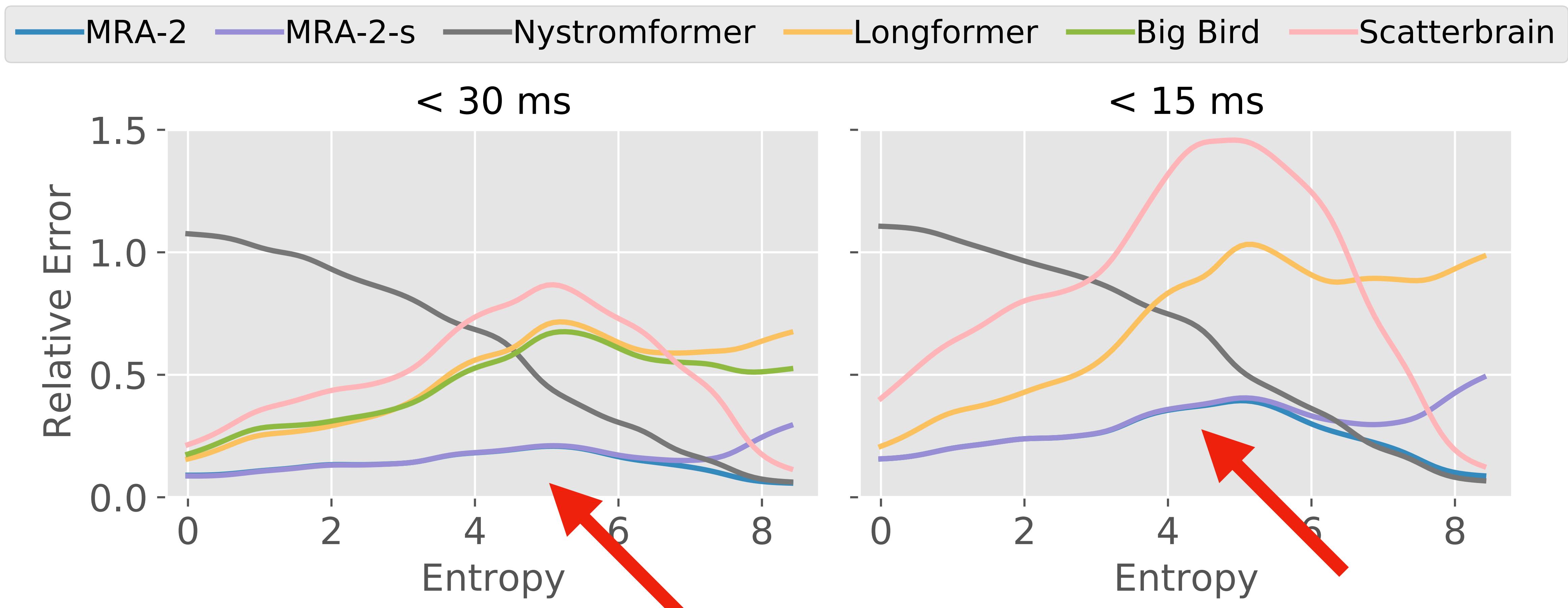
Evaluations

Entropy v.s. Approximation



Evaluations

Entropy v.s. Approximation

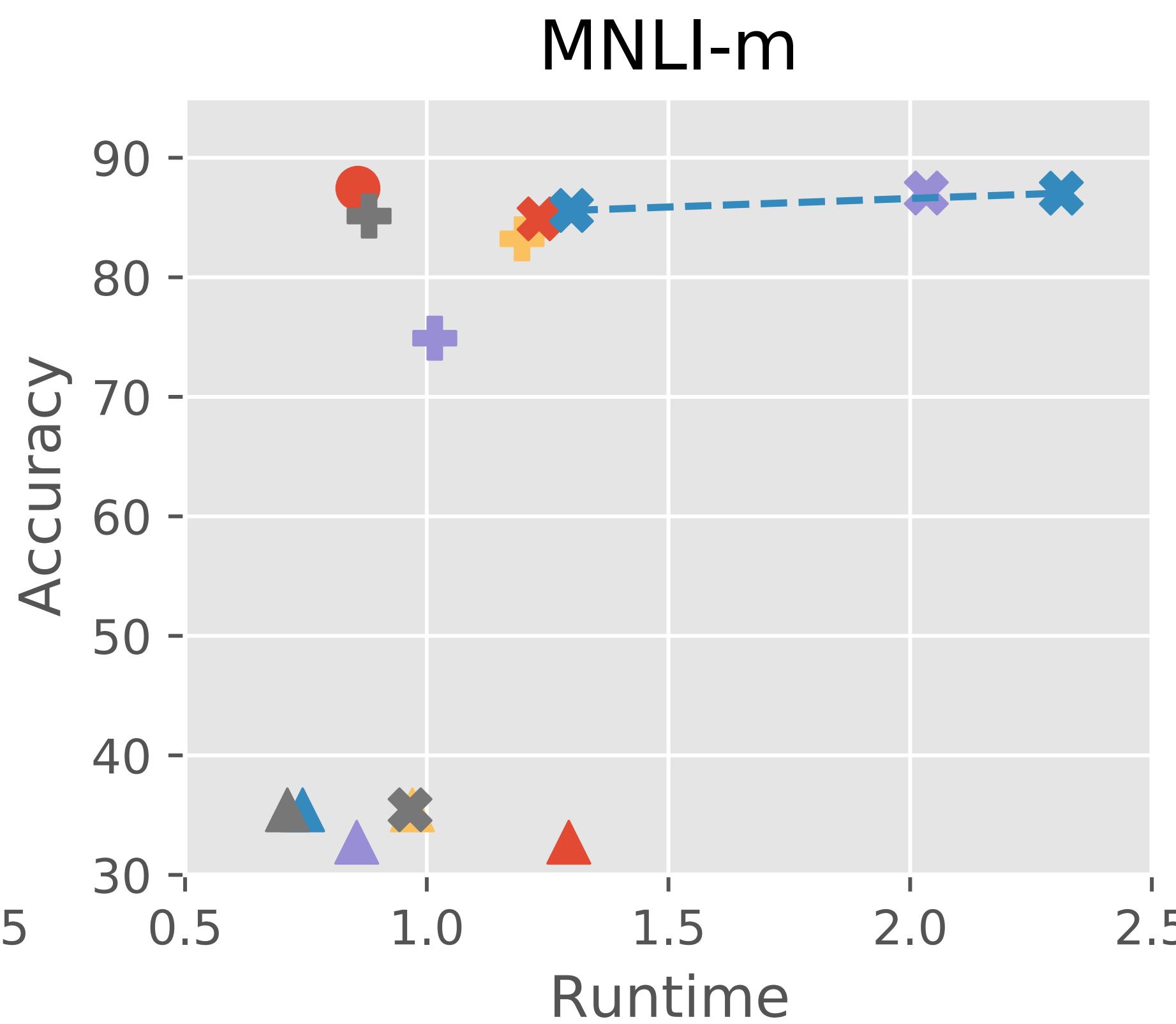
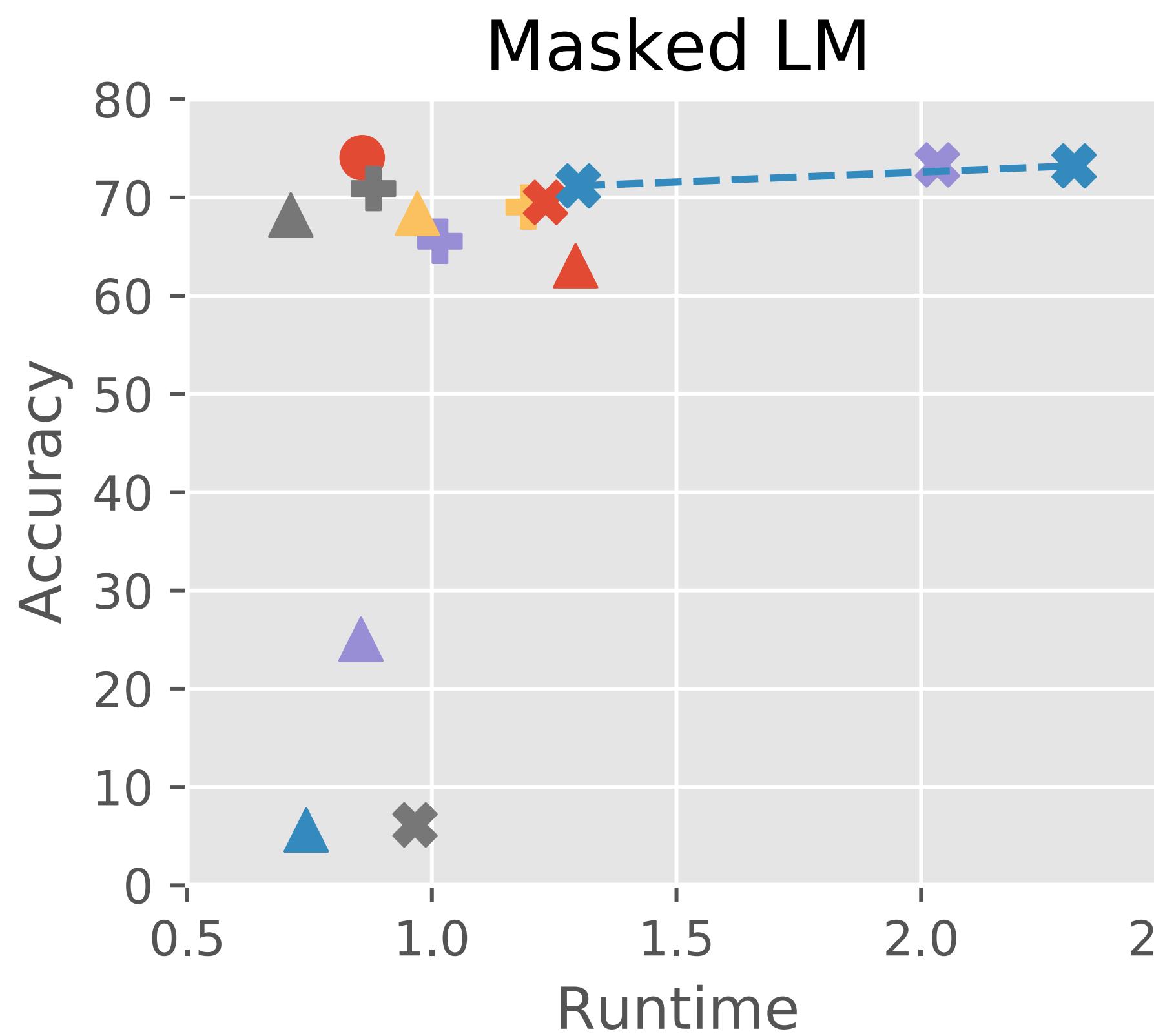


Evaluations

RoBERTa-base-512

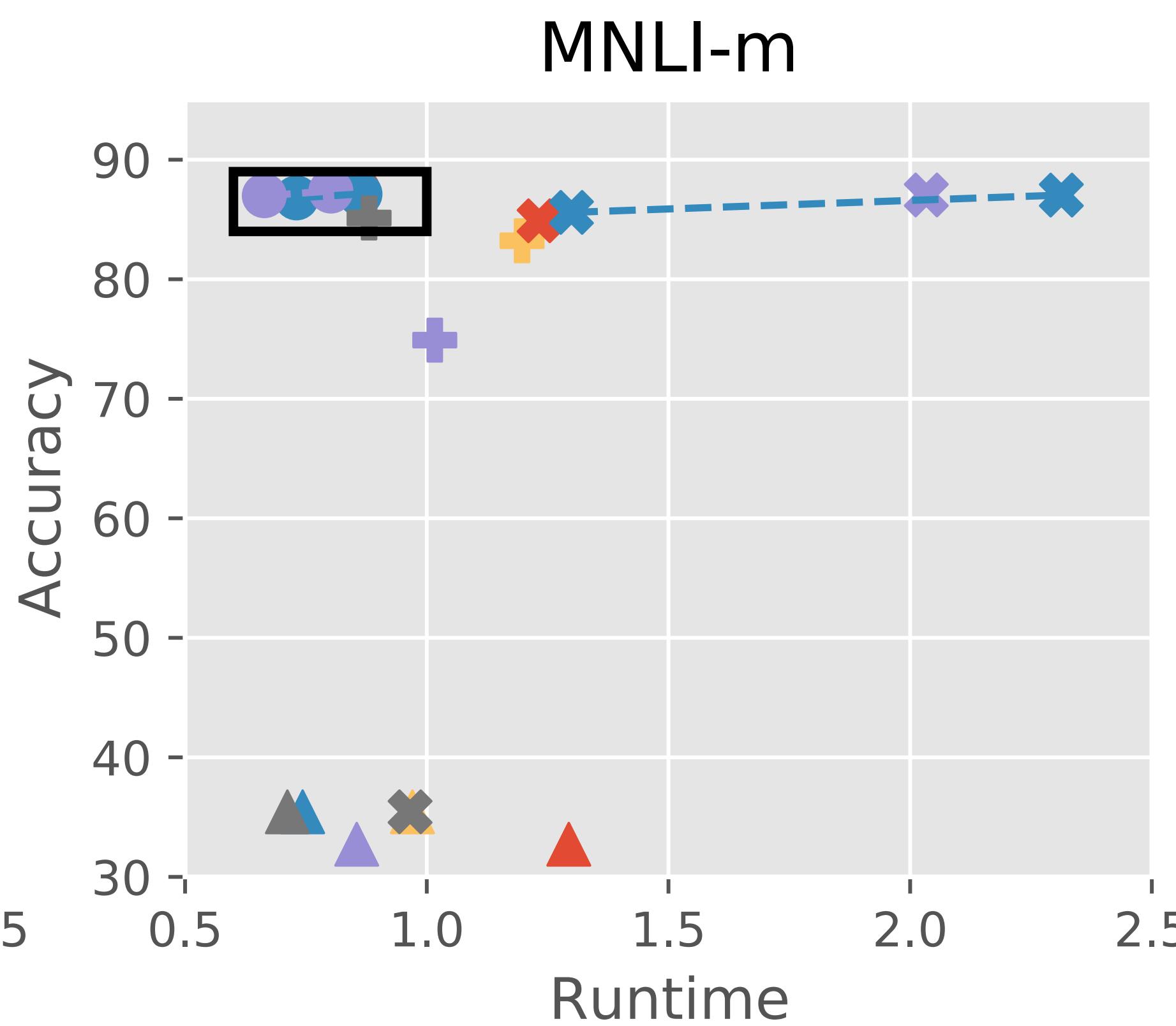
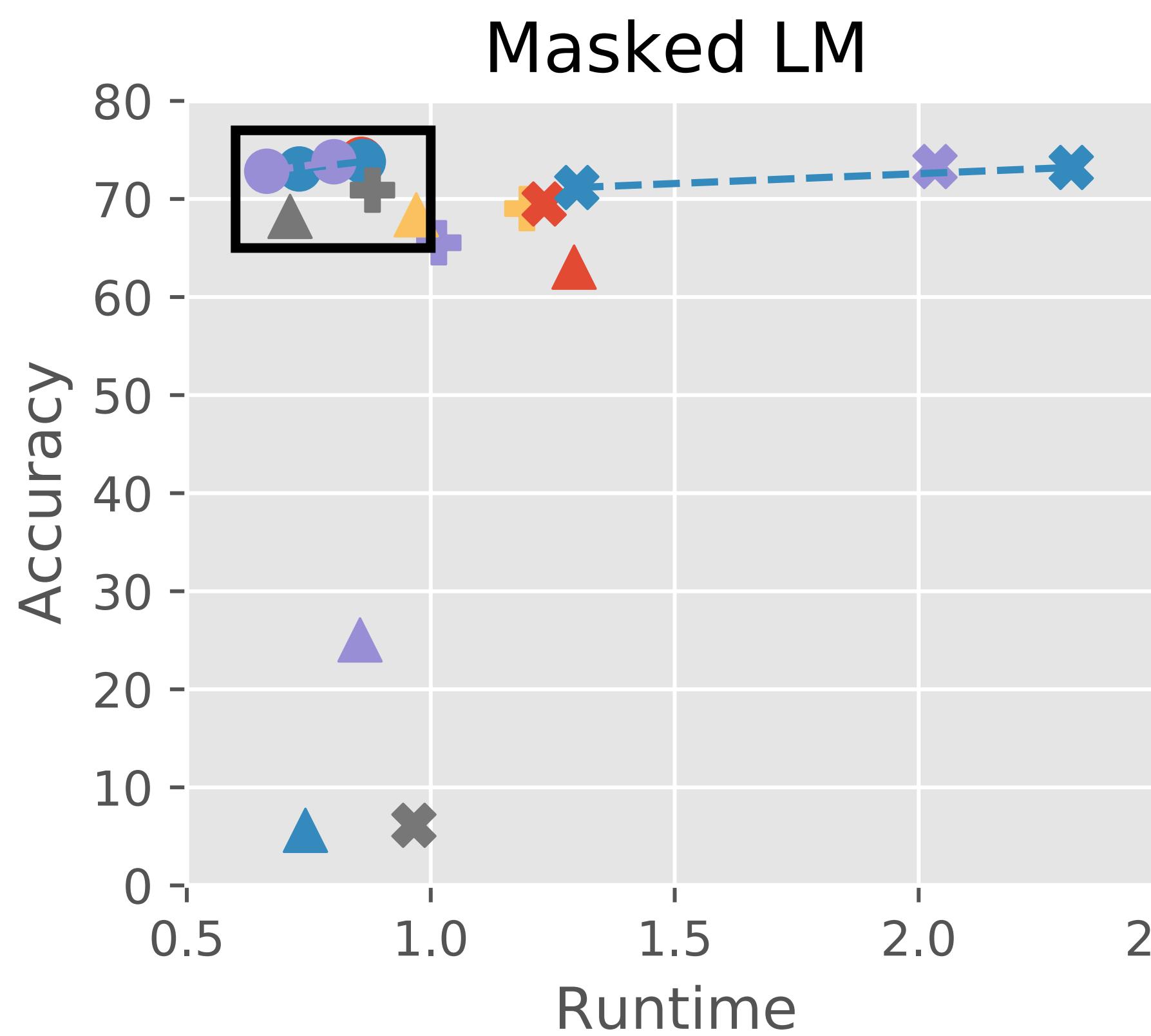
Evaluations

RoBERTa-base-512



Evaluations

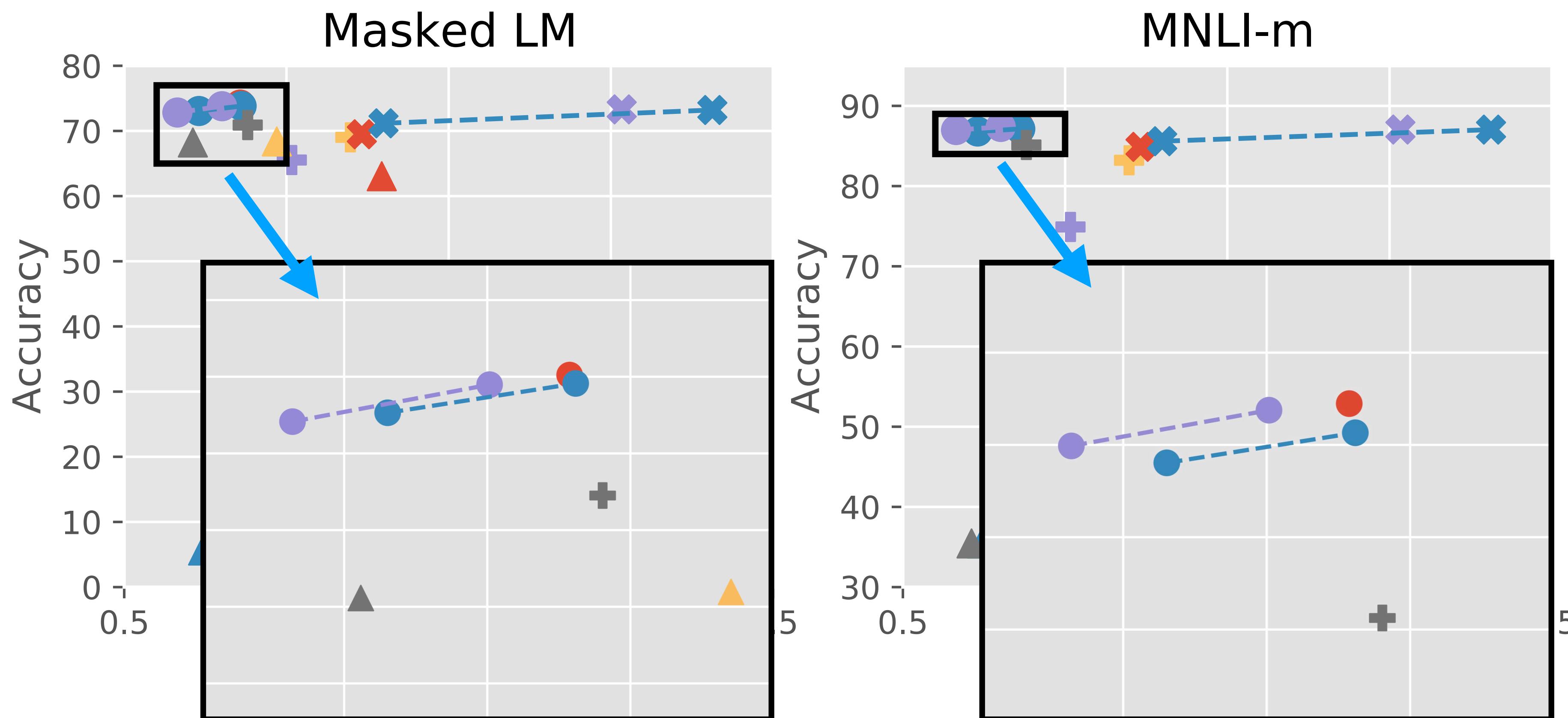
RoBERTa-base-512



- Transformer
- MRA-2
- MRA-2-s
- Performer
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

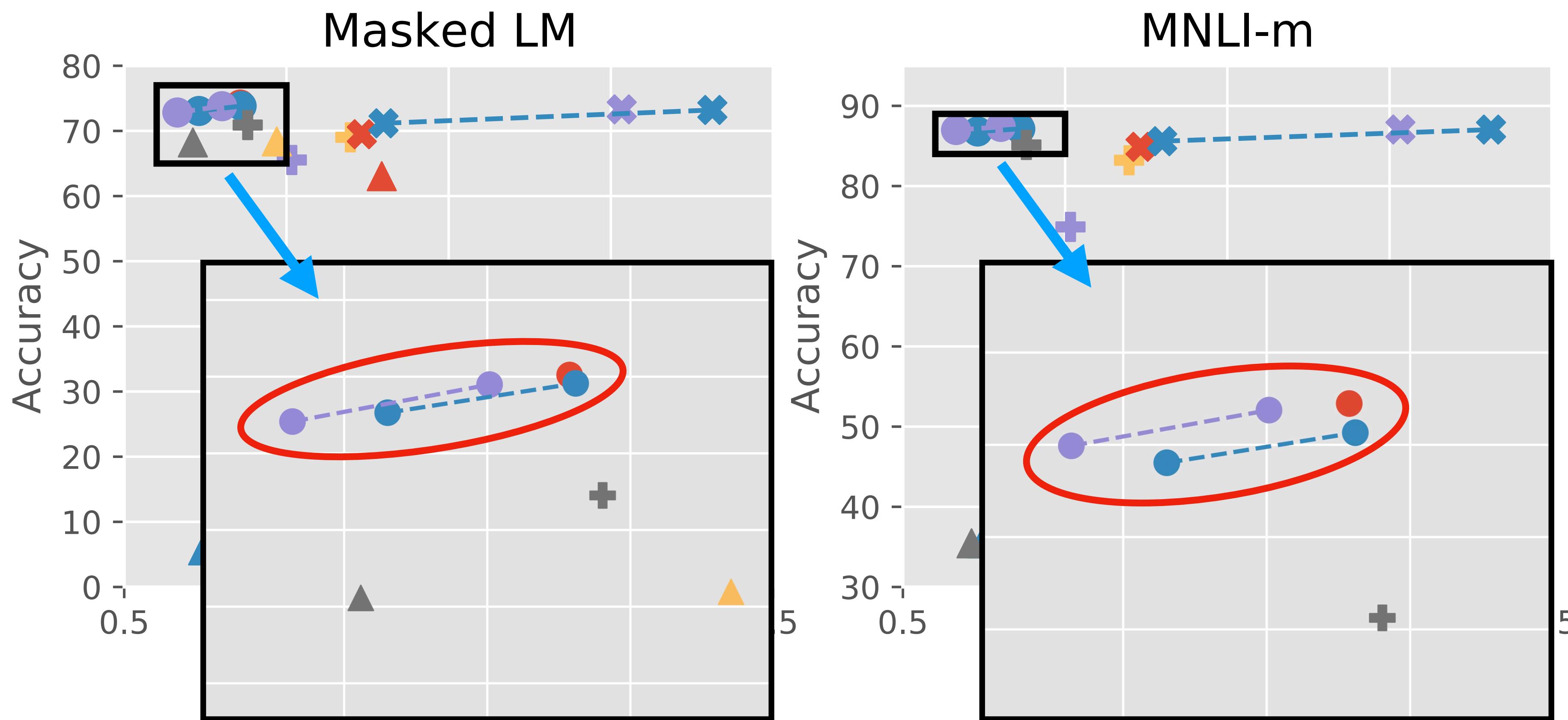
RoBERTa-base-512



- Transformer
- MRA-2
- MRA-2-s
- Performer
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

RoBERTa-base-512



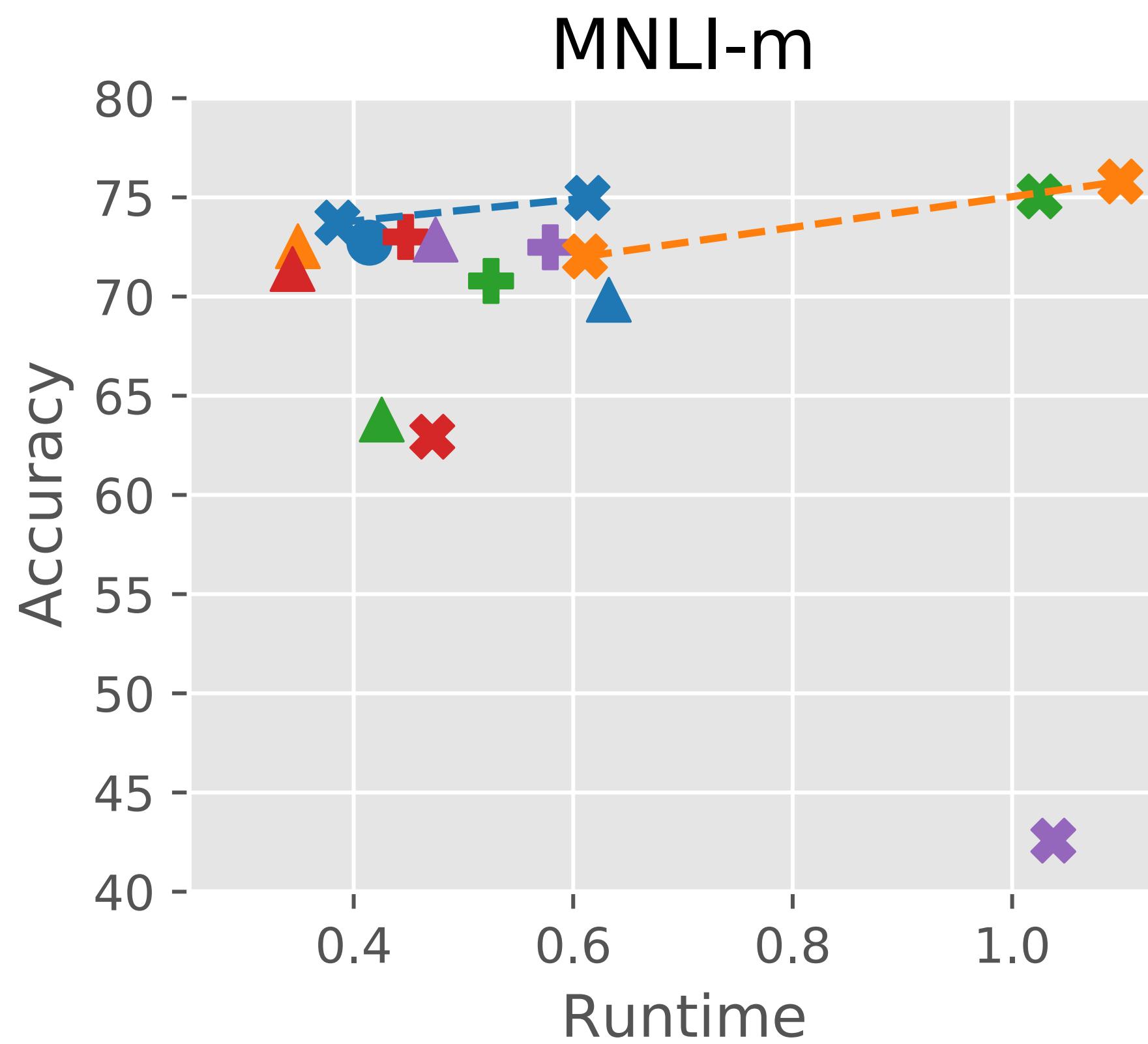
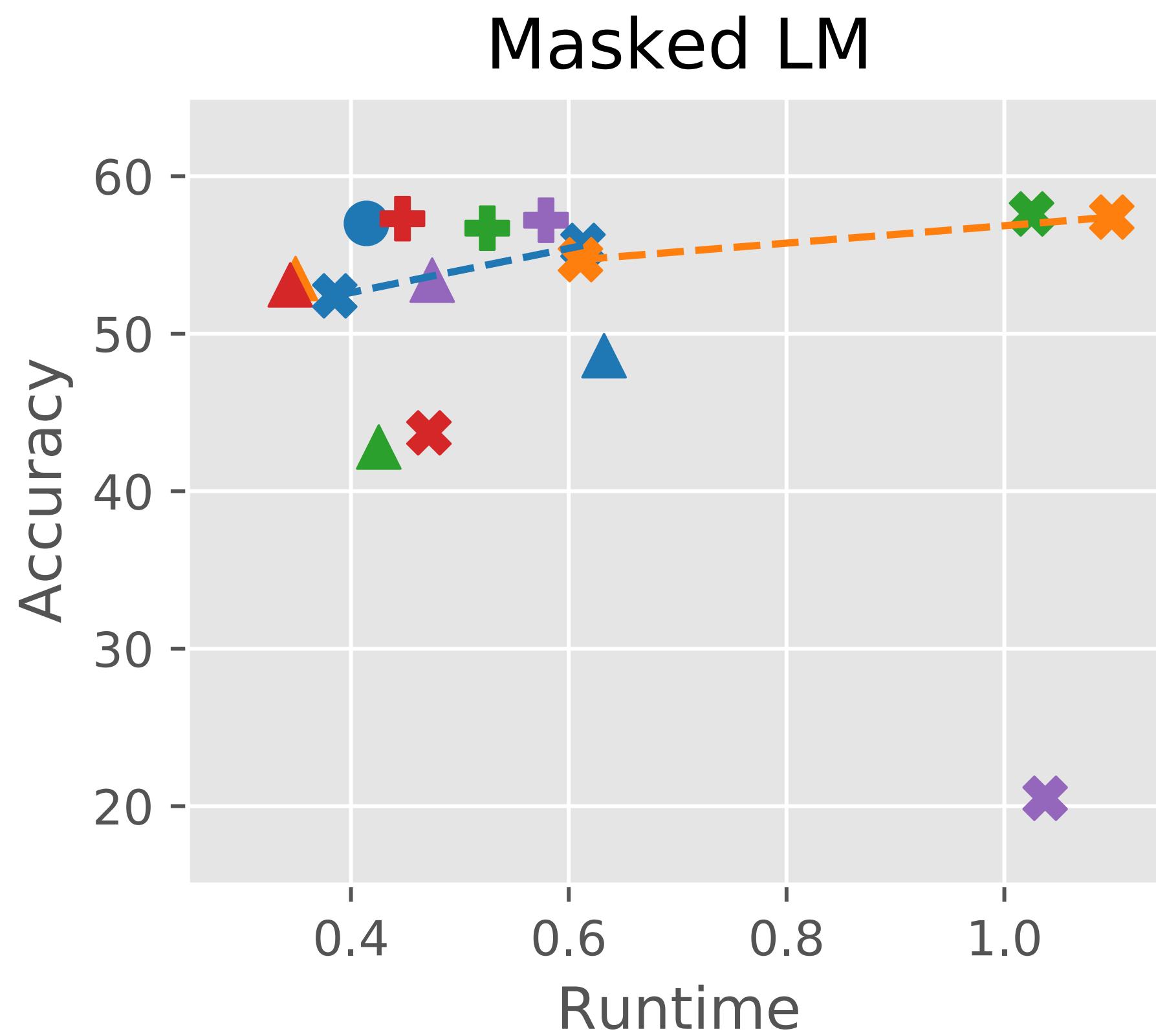
- Transformer
- MRA-2
- MRA-2-s
- Performer
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

RoBERTa-small-512

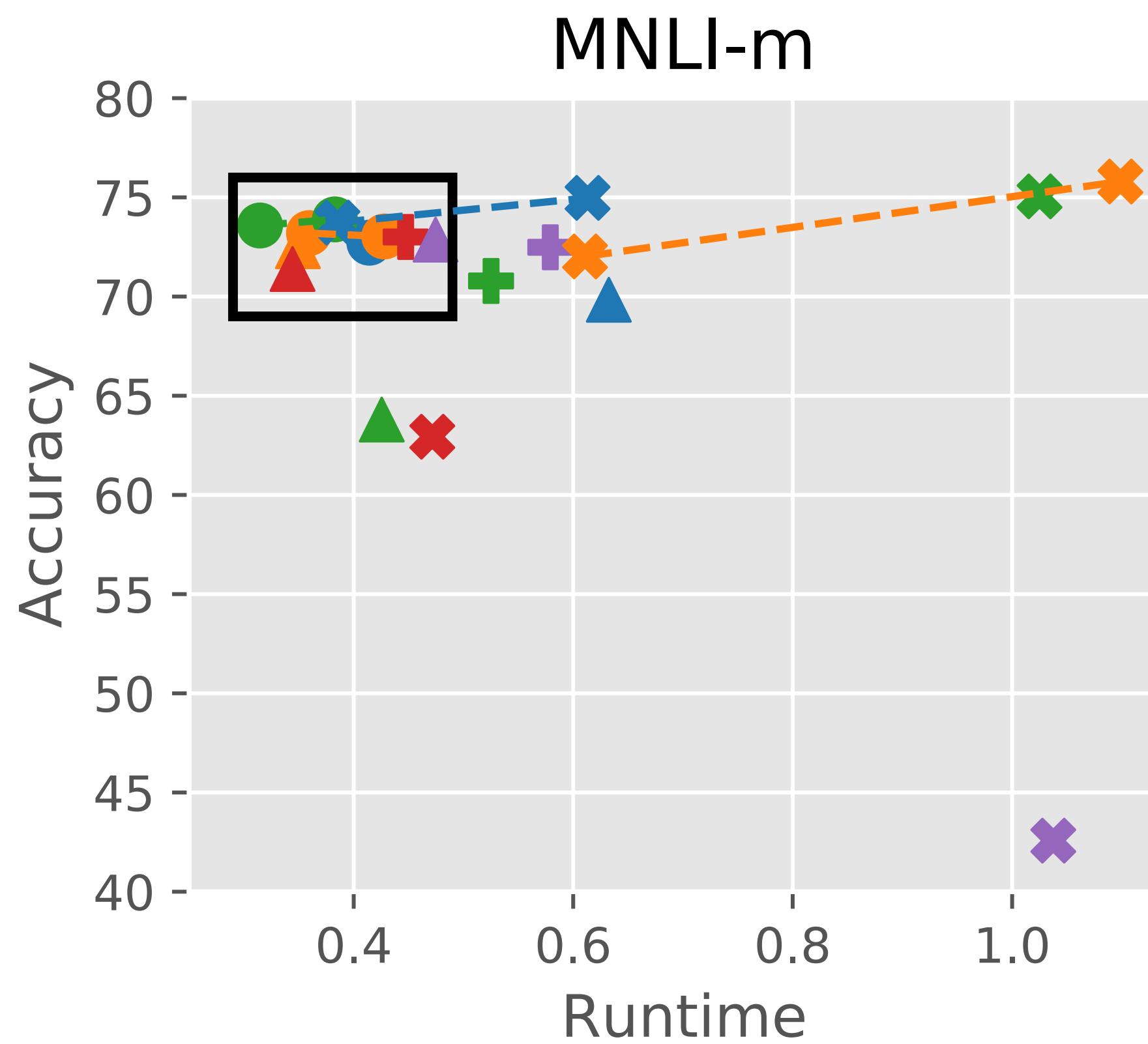
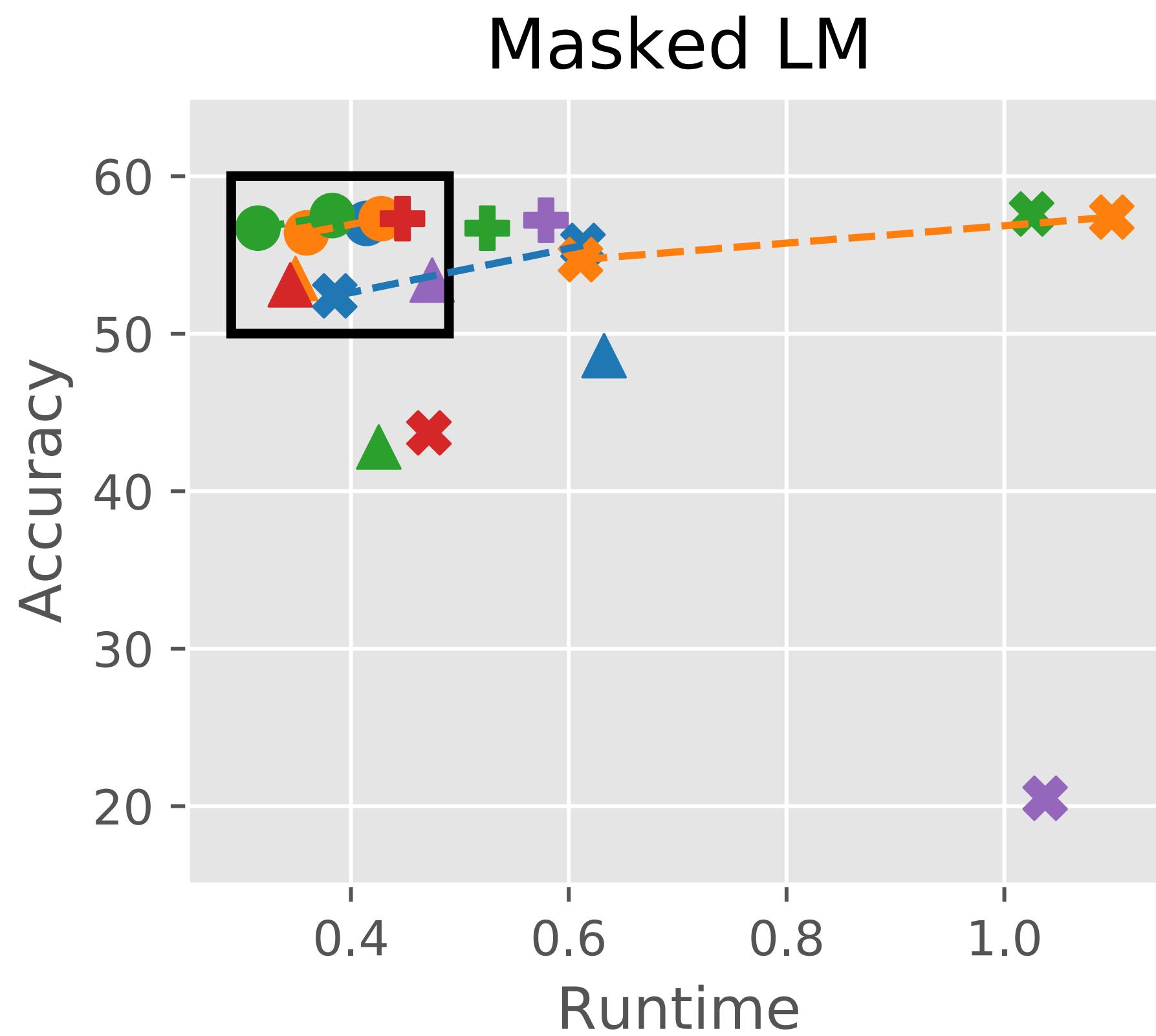
Evaluations

RoBERTa-small-512



Evaluations

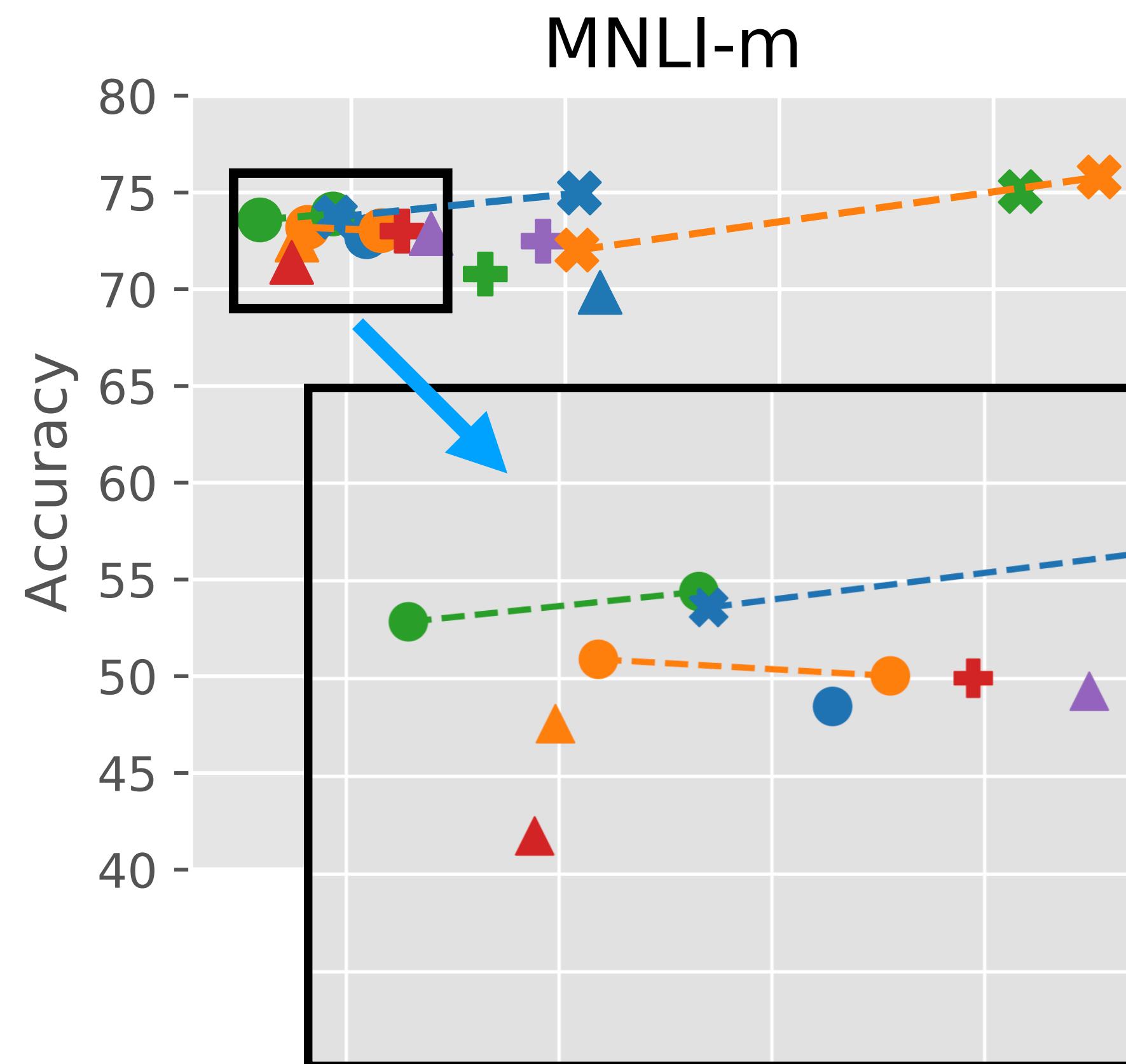
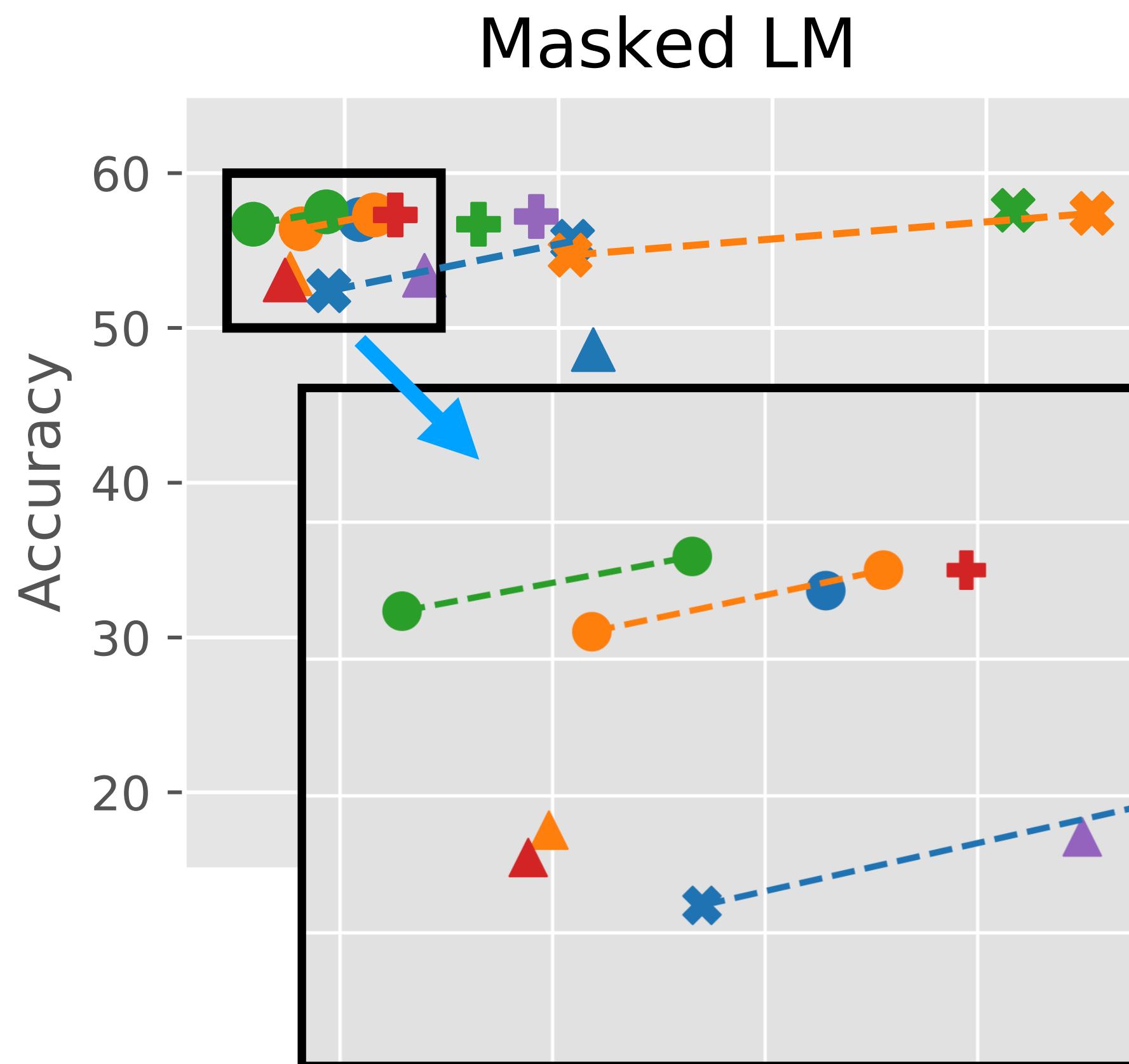
RoBERTa-small-512



- Transformer
- MRA-2
- MRA-2-s
- Performers
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

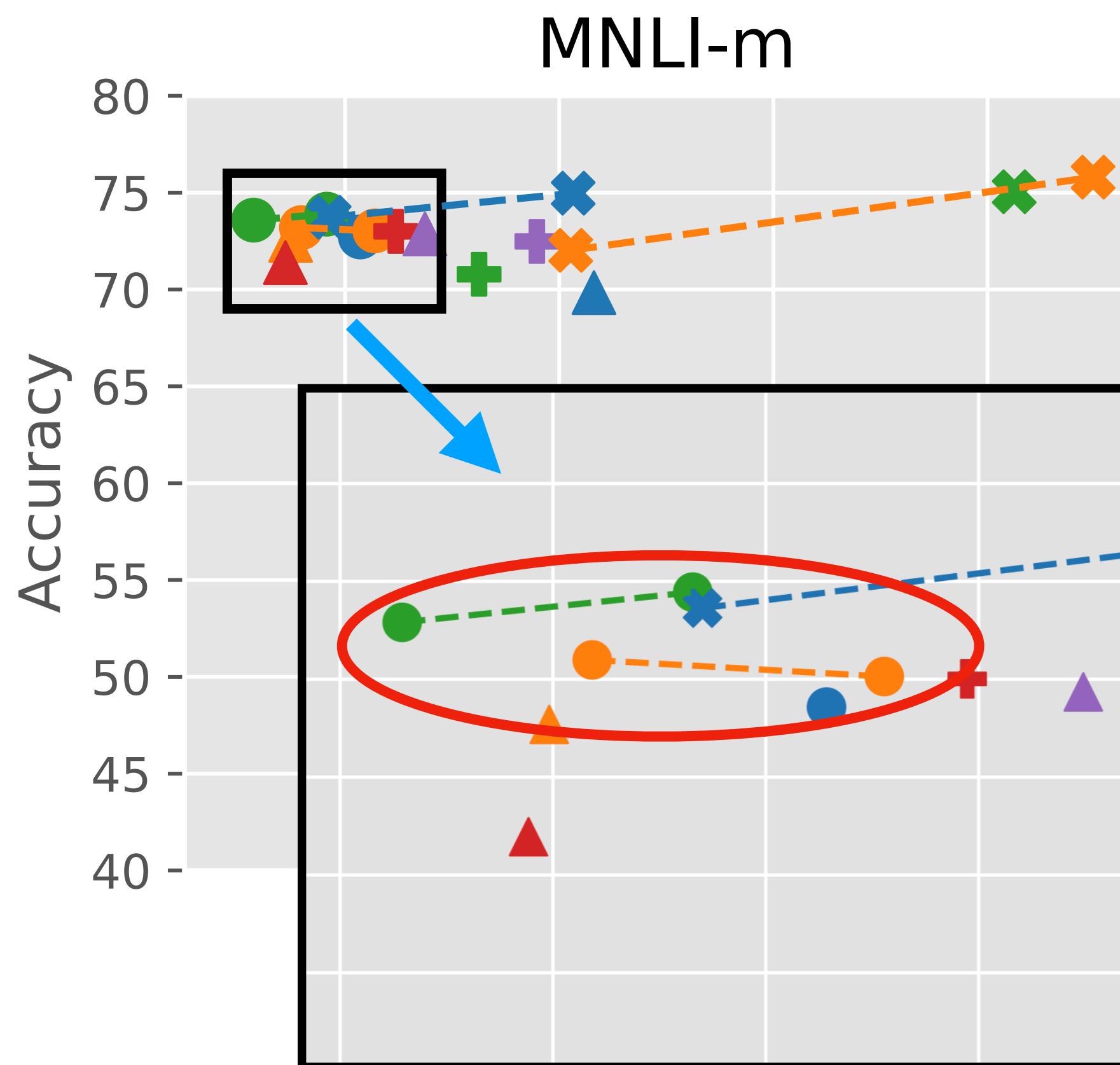
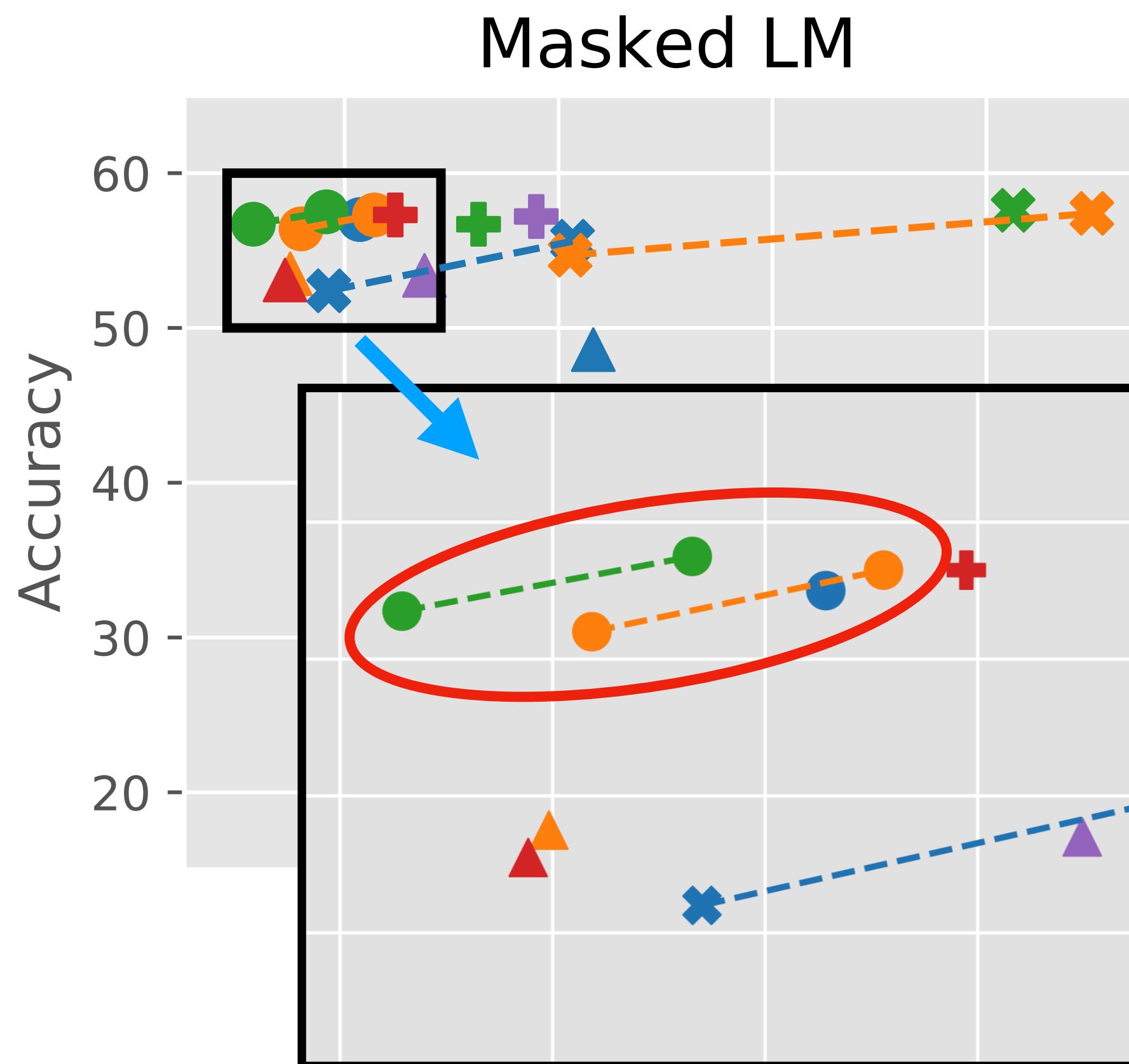
RoBERTa-small-512



- Transformer
- MRA-2
- MRA-2-s
- Performers
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

RoBERTa-small-512



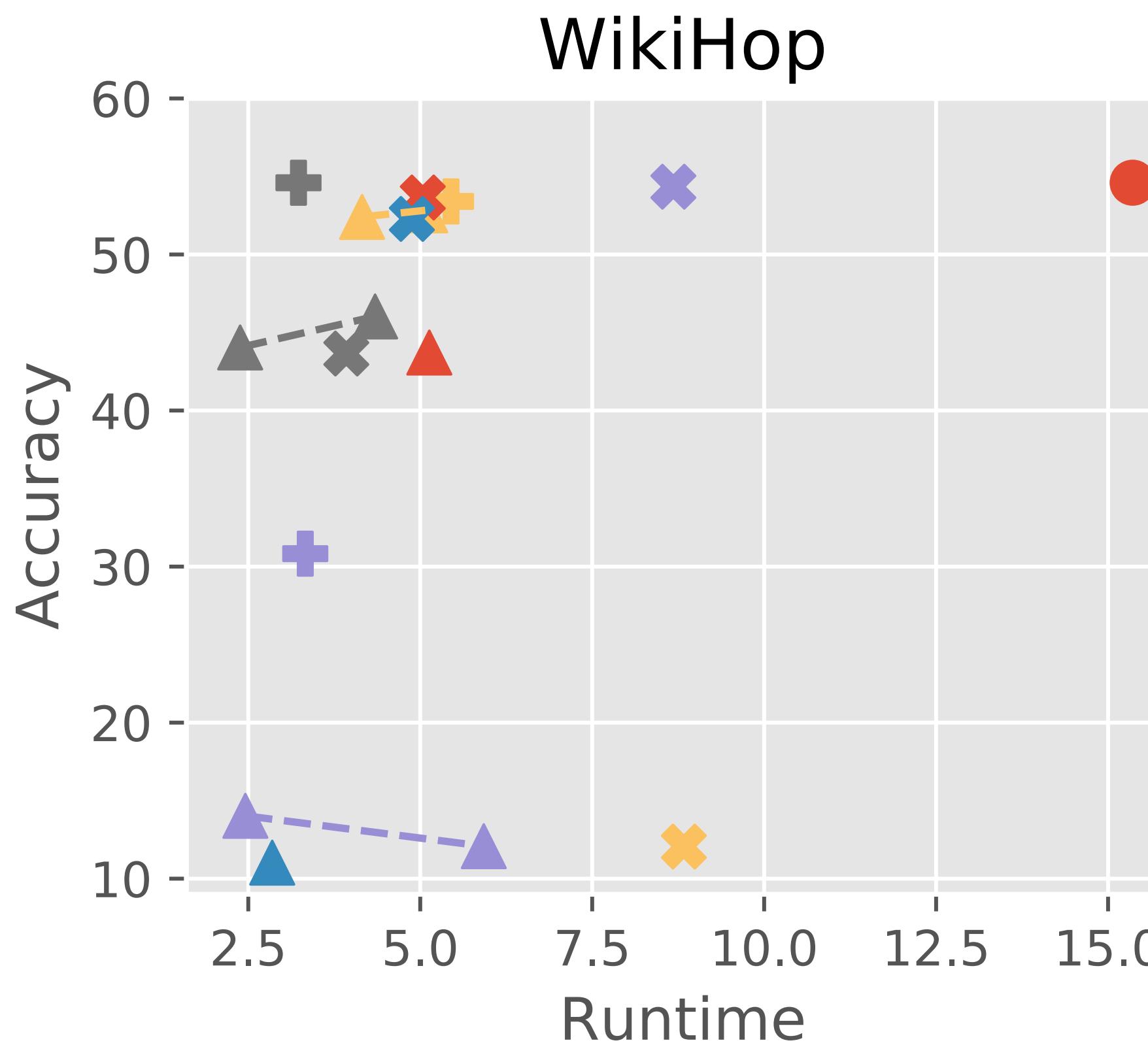
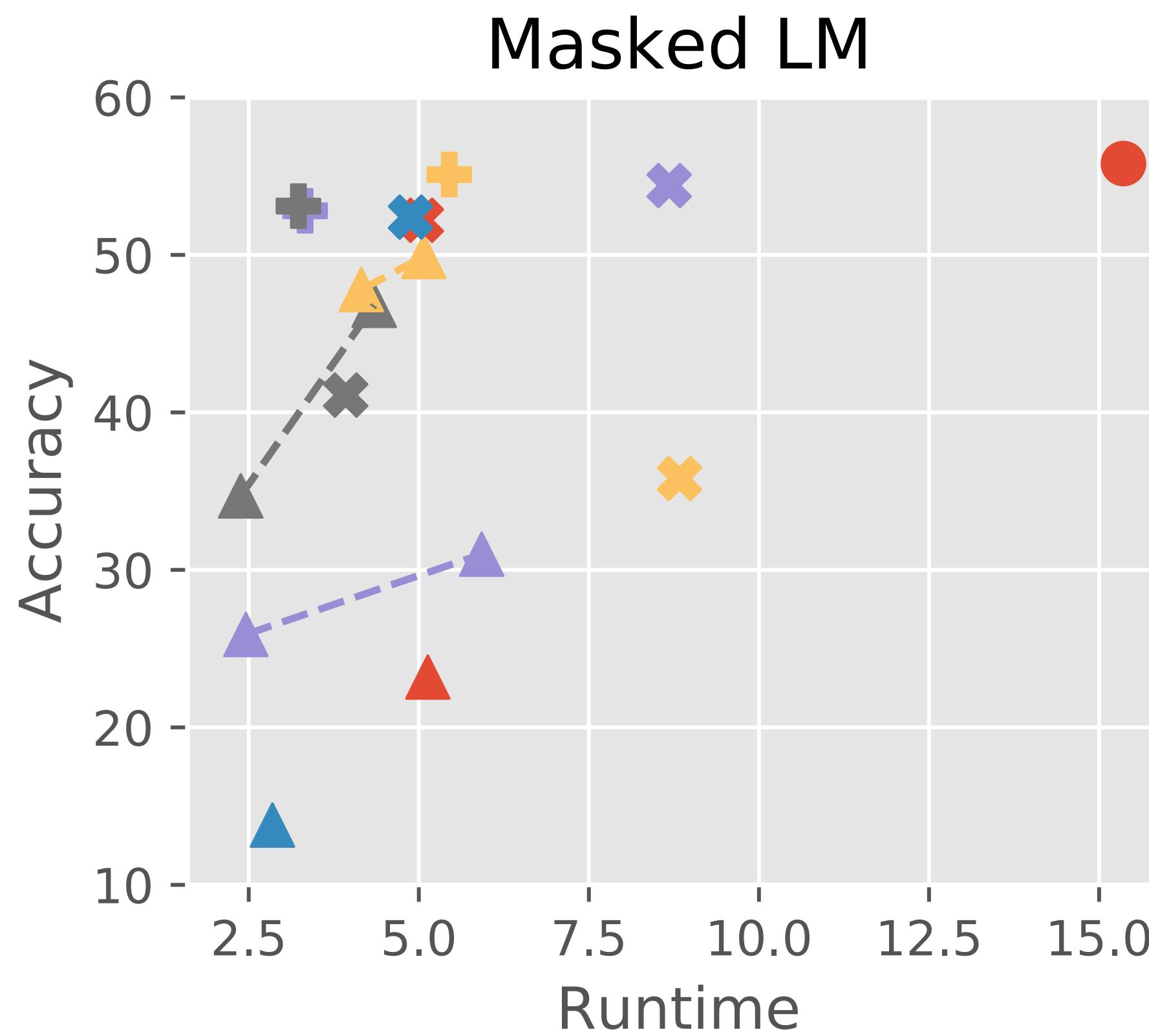
- Transformer
- MRA-2
- MRA-2-s
- ▲ Performer
- ▲ Linformer
- ▲ SOFT
- ▲ SOFT + Conv
- ▲ Nystromformer
- ▲ Nystrom + Conv
- ▲ YOZO
- ▲ YOZO + Conv
- ▲ Reformer
- ▲ Longformer
- ▲ Big Bird
- ▲ H-Transformer-1D
- ▲ Scatterbrain

Evaluations

RoBERTa-small-4096

Evaluations

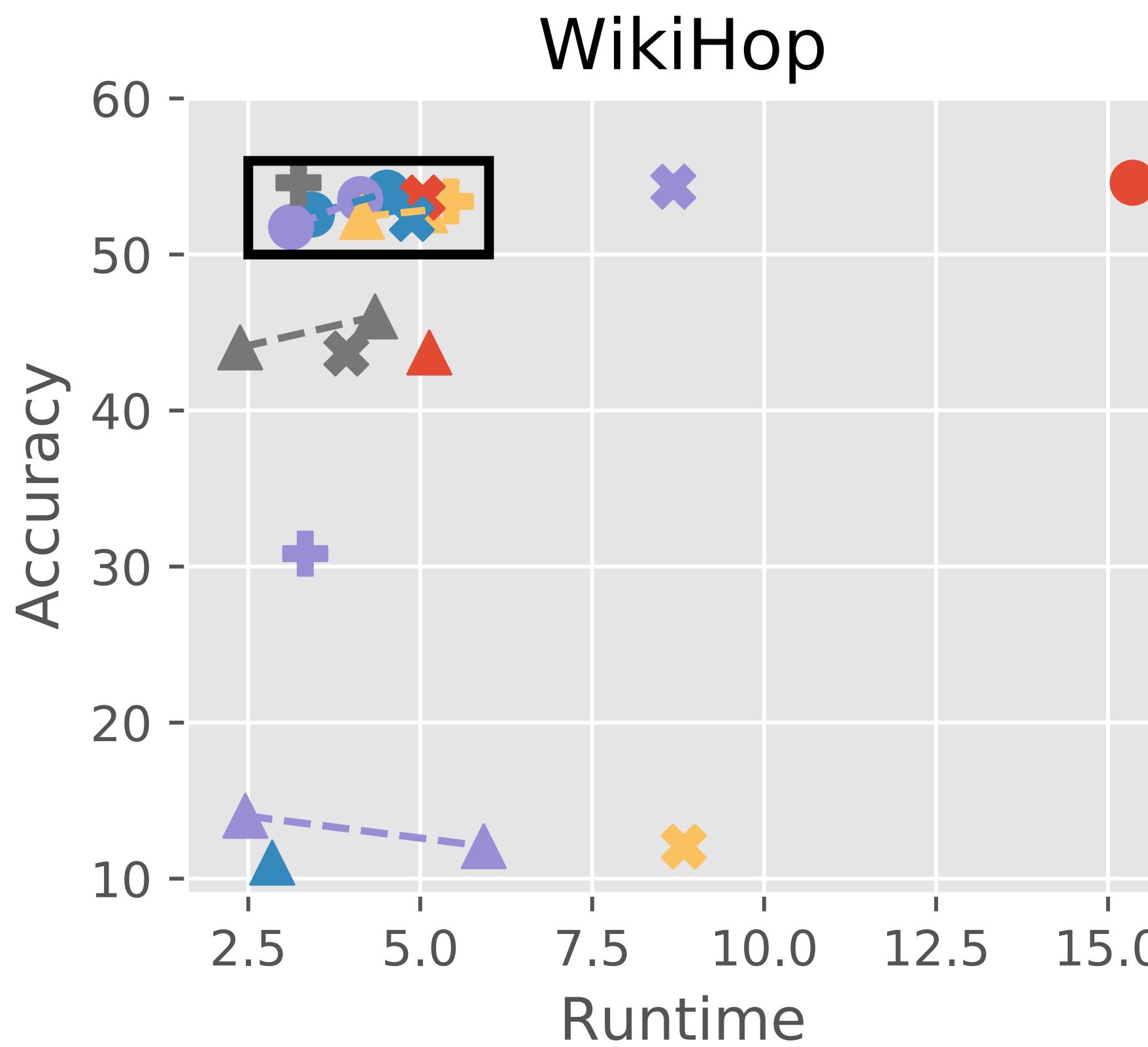
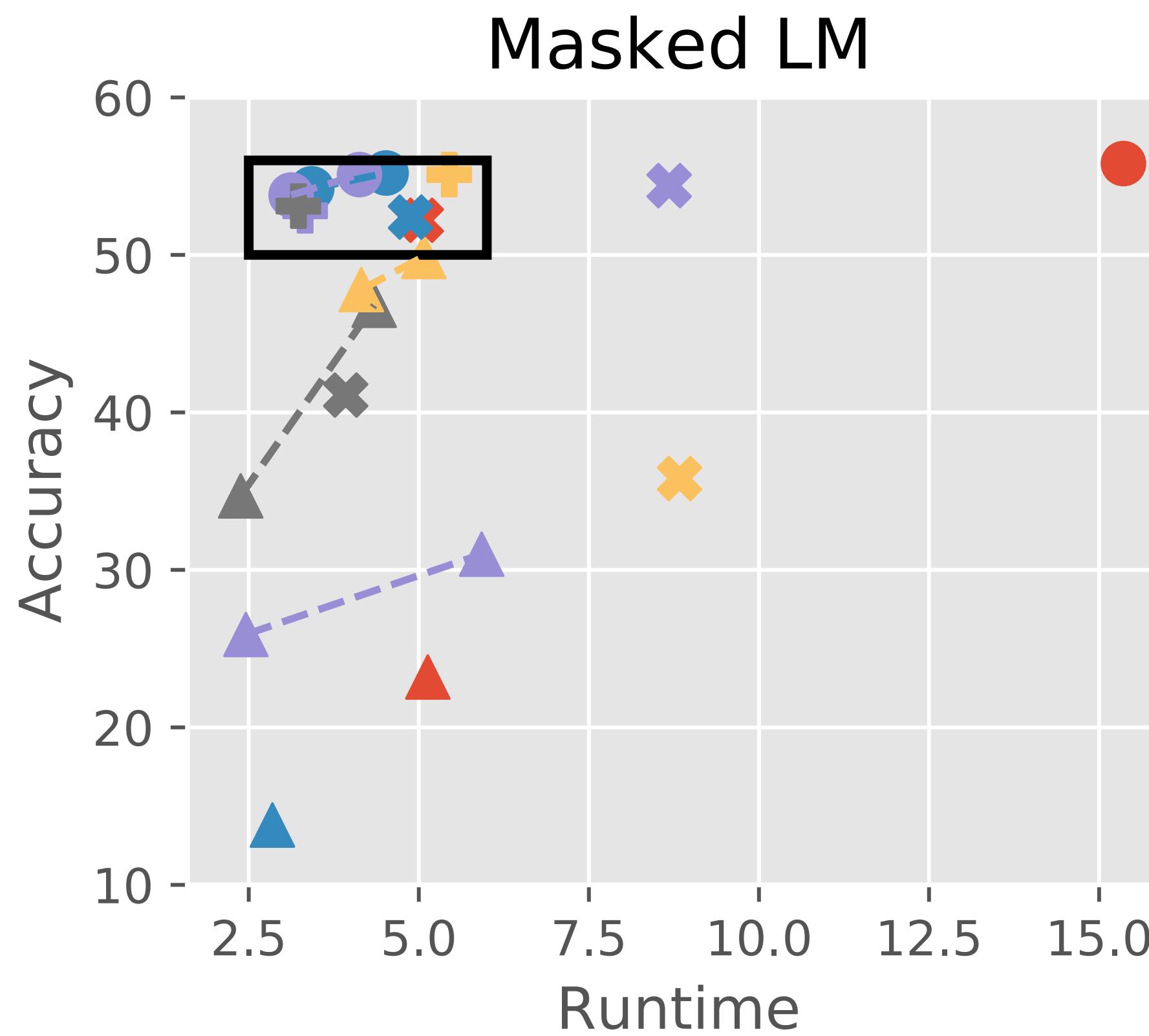
RoBERTa-small-4096



- Transformer
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YOSO
- YOSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

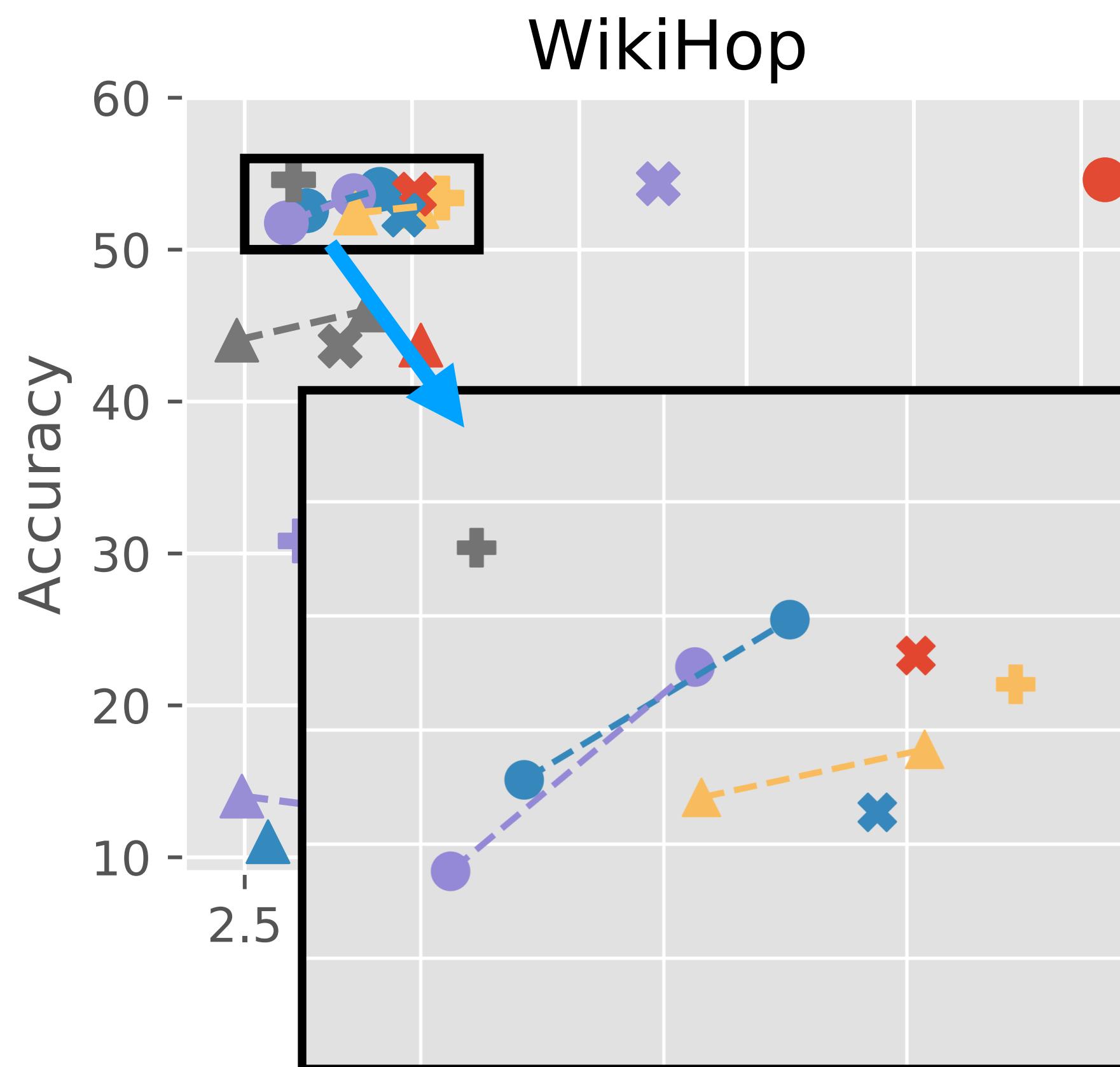
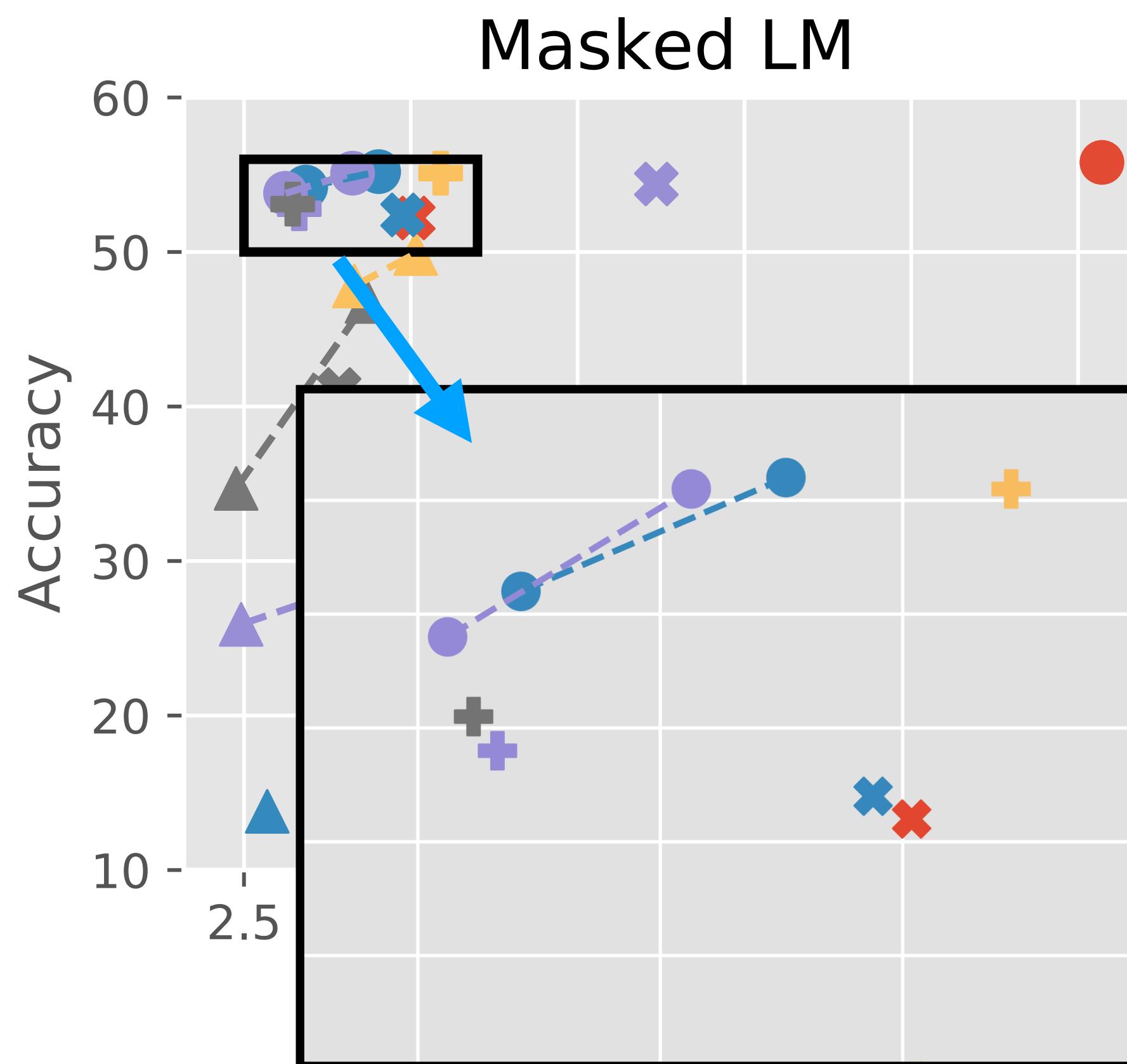
RoBERTa-small-4096



- Transformer
- MRA-2
- MRA-2-s
- Performer
- Linformer
- SOFT
- YO SO + Conv
- Nystromformer
- Nystrom + Conv
- YO SO
- YO SO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

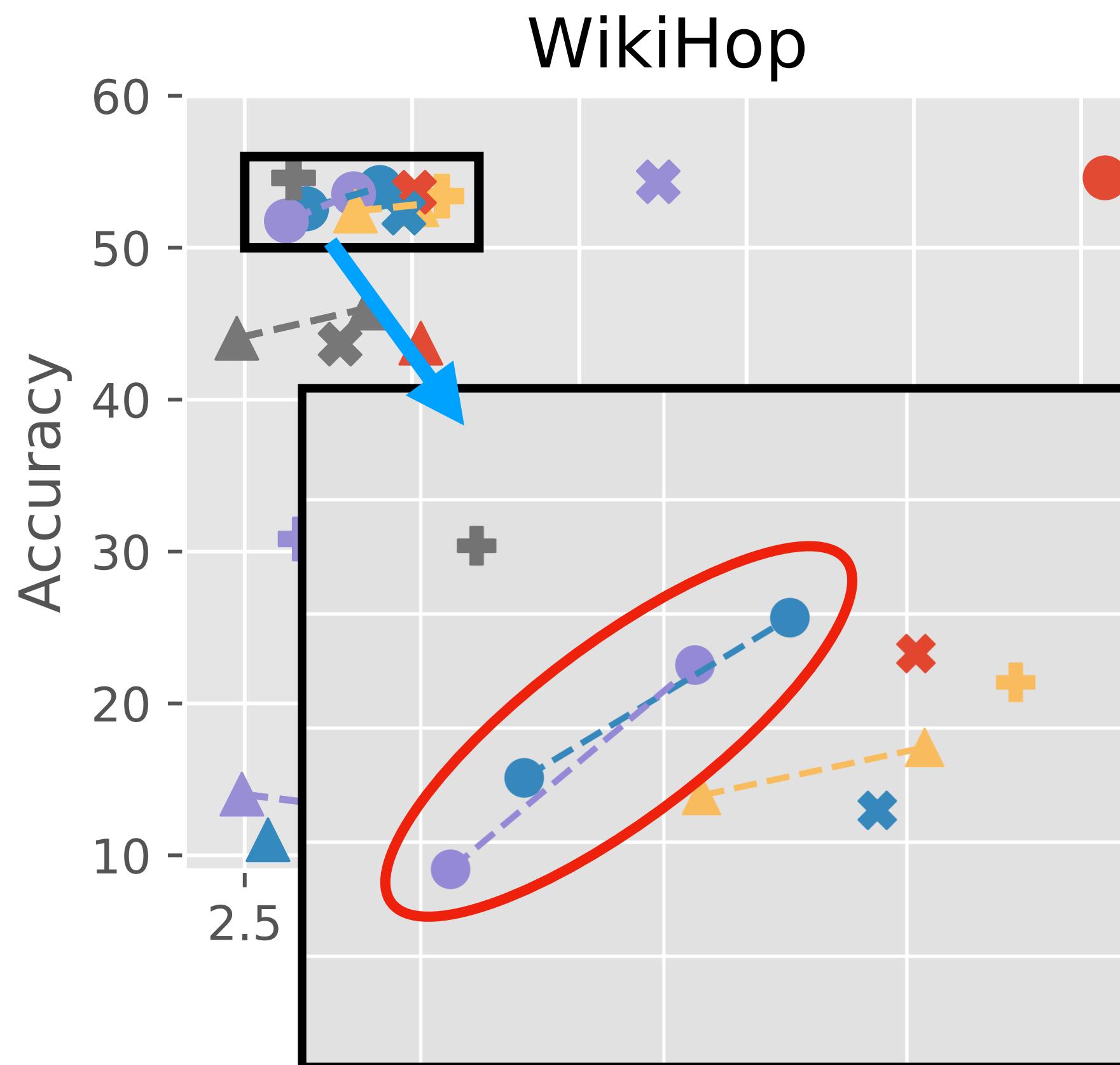
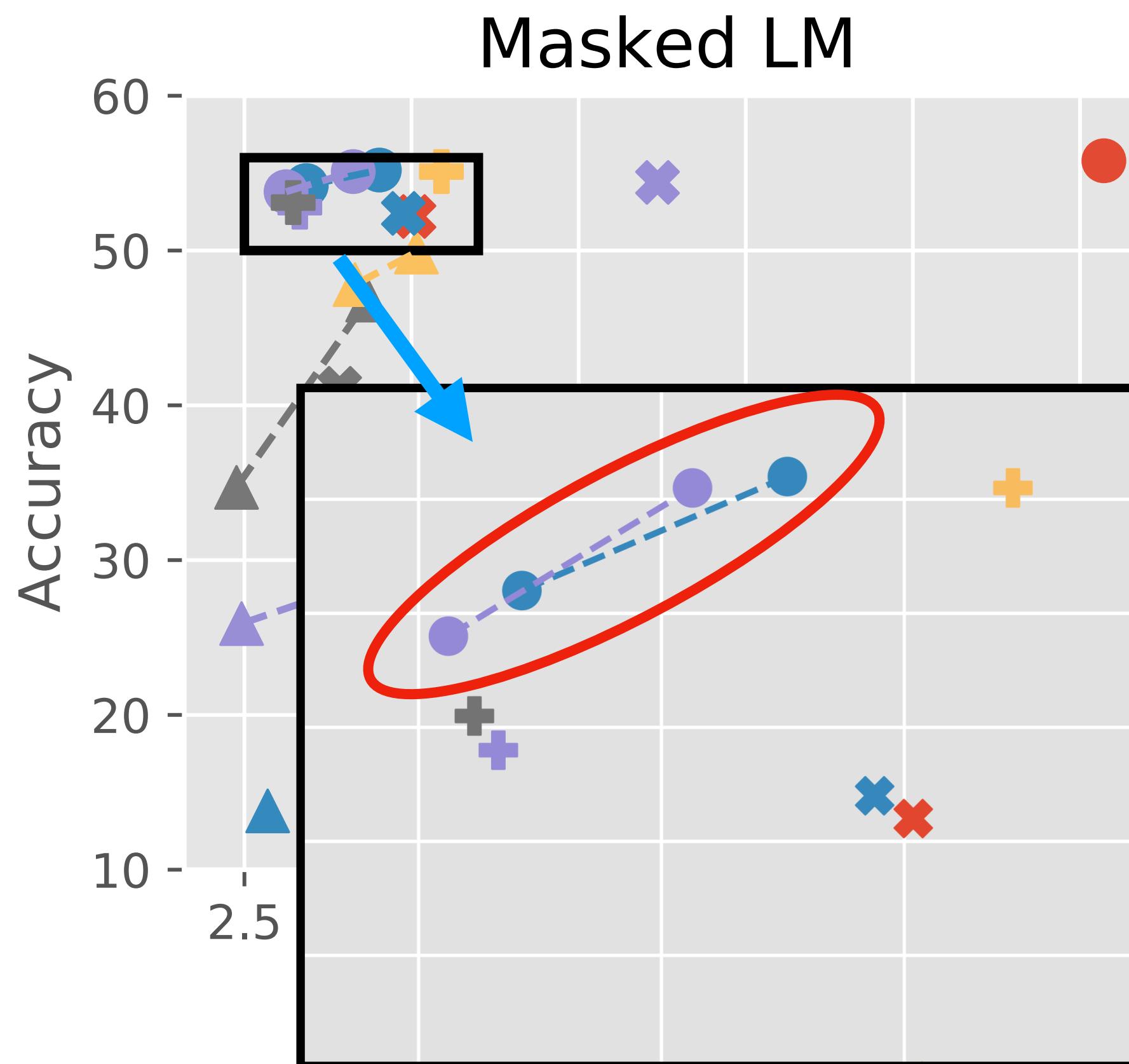
RoBERTa-small-4096



- Transformer
- MRA-2
- MRA-2-s
- Performers
- Linformer
- SOFT
- YOSO
- YOSO + Conv
- Nystromformer
- Nystrom + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

RoBERTa-small-4096



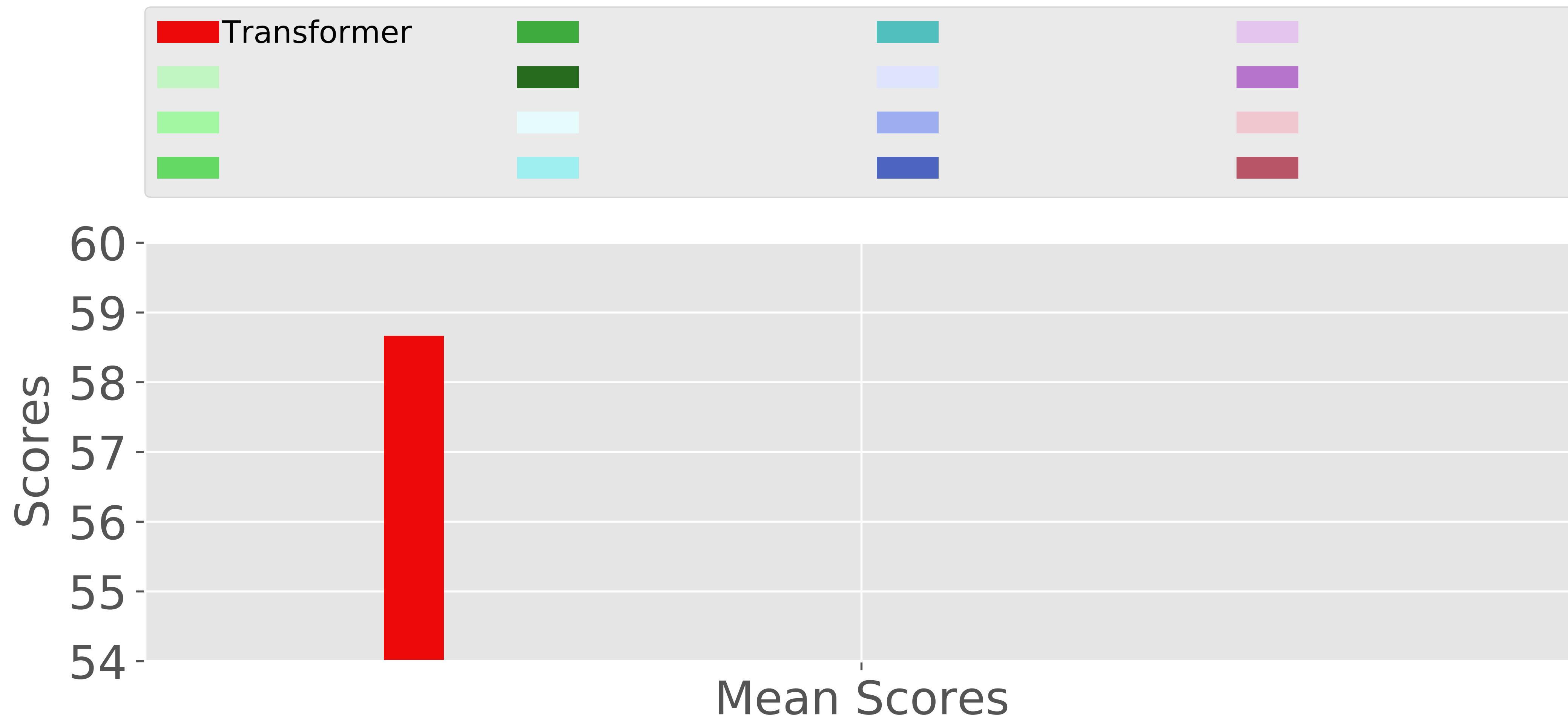
- Transformer
- MRA-2
- MRA-2-s
- Performer
- Linformer
- SOFT
- SOFT + Conv
- Nystromformer
- Nystrom + Conv
- YSO
- YSO + Conv
- Reformer
- Longformer
- Big Bird
- H-Transformer-1D
- Scatterbrain

Evaluations

Long Range Arena Benchmark

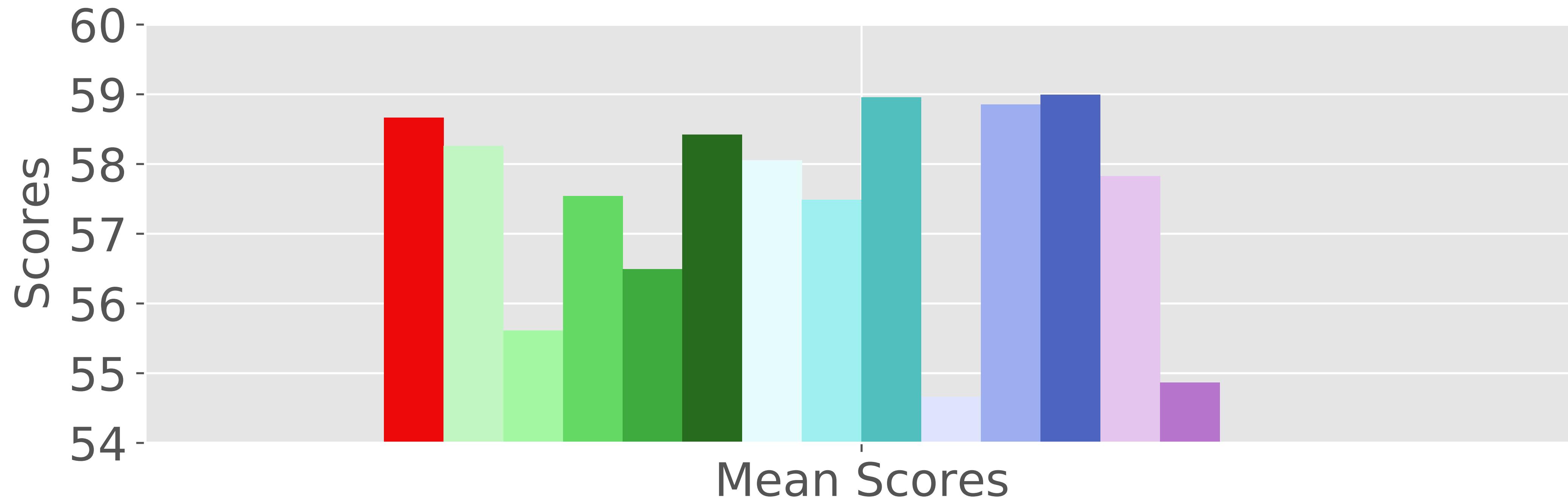
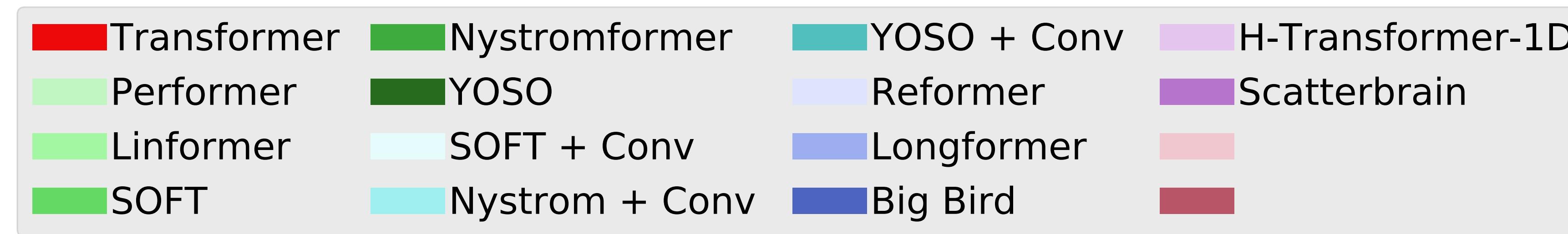
Evaluations

Long Range Arena Benchmark



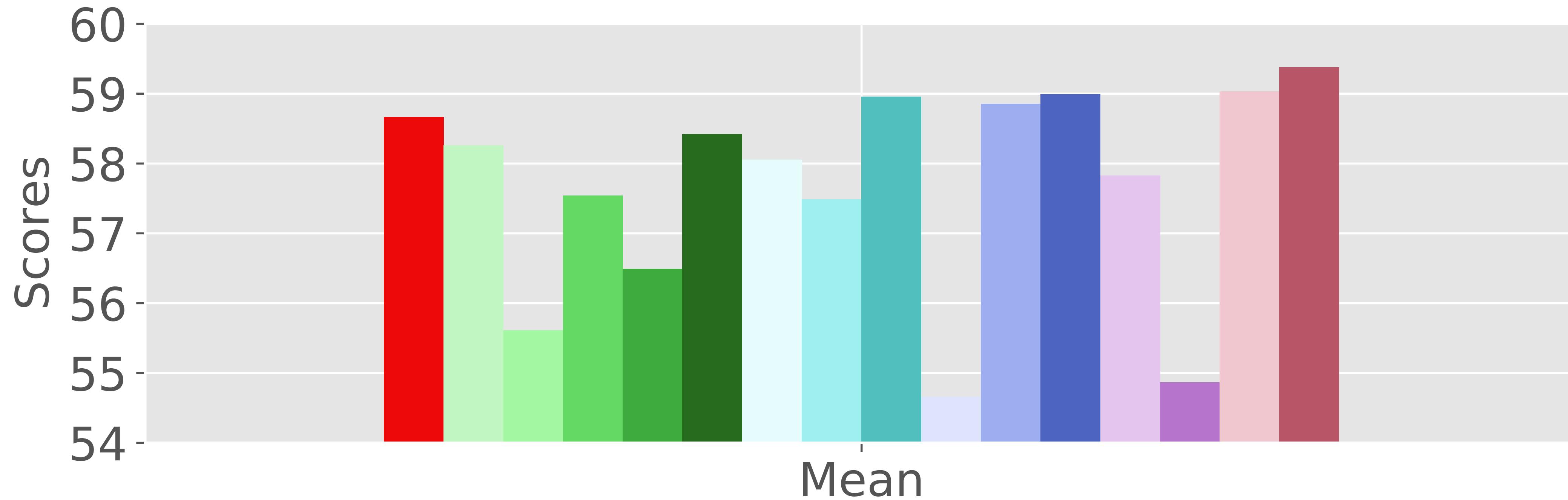
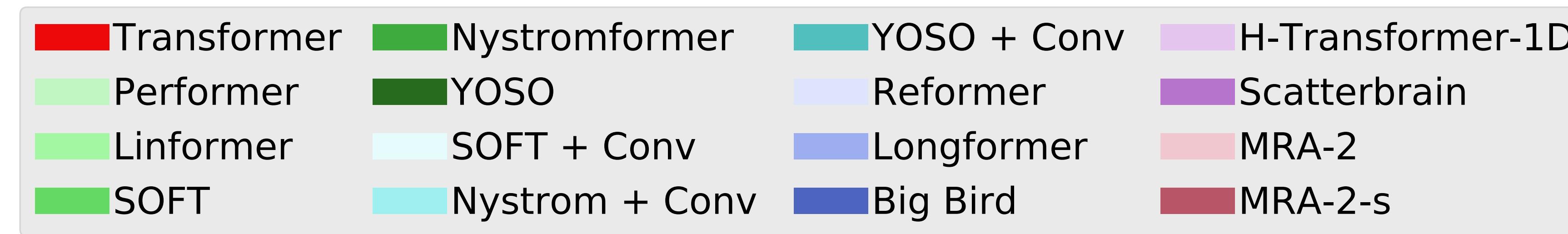
Evaluations

Long Range Arena Benchmark



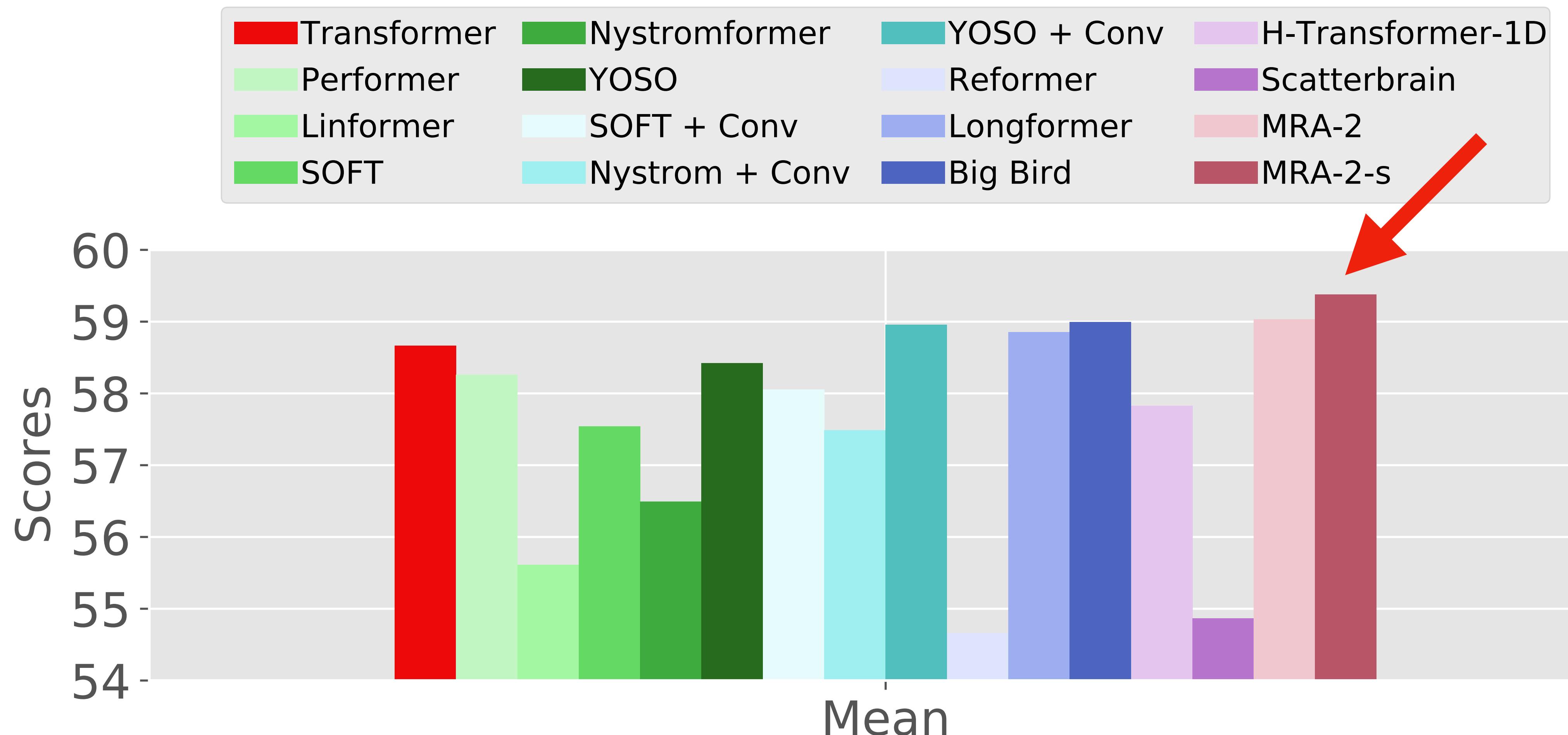
Evaluations

Long Range Arena Benchmark



Evaluations

Long Range Arena Benchmark

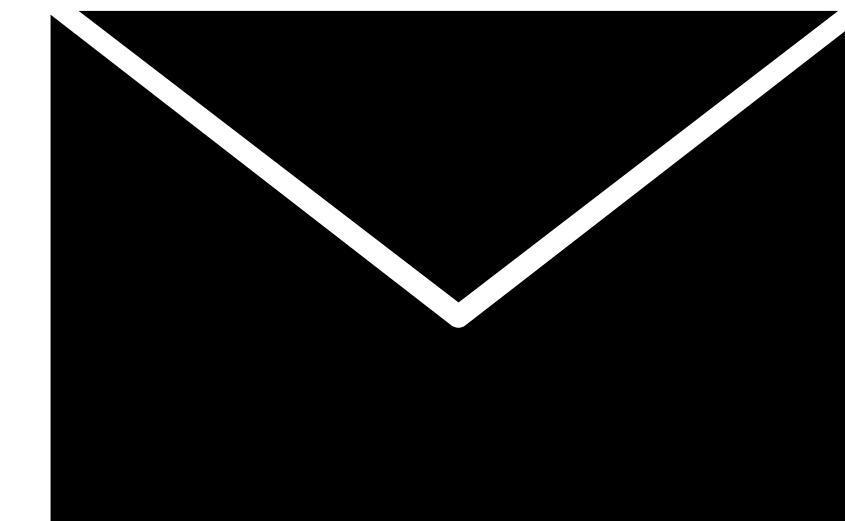


Takeaway

- Multi-resolution analysis (MRA) ideas offer a win-win for efficient self-attention
 - potential access to a rich MRA theory
 - immediate practical benefits on efficiency and accuracy



[mlpen/mra-attention](https://github.com/mlpen/mra-attention)



zzeng38@wisc.edu

mlpen / mra-attention Public

Pin Unwatch 1 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

mlpen update 5fc2033 21 minutes ago 4 commits

ImageNet	update	21 minutes ago
LRA	update	21 minutes ago
RoBERTa	add files	25 minutes ago
supplement_code	add files	25 minutes ago
.DS_Store	update	21 minutes ago
README.md	Update README.md	23 minutes ago

README.md

Official Repo for Multi Resolution Analysis (MRA) for Approximate Self-Attention

About

No description, website, or topics provided.

Readme 0 stars 1 watching 0 forks

Releases

No releases published [Create a new release](#)

Packages

No packages published [Publish your first package](#)

<https://github.com/mlpen/mra-attention>