Green AI

Roy Schwartz

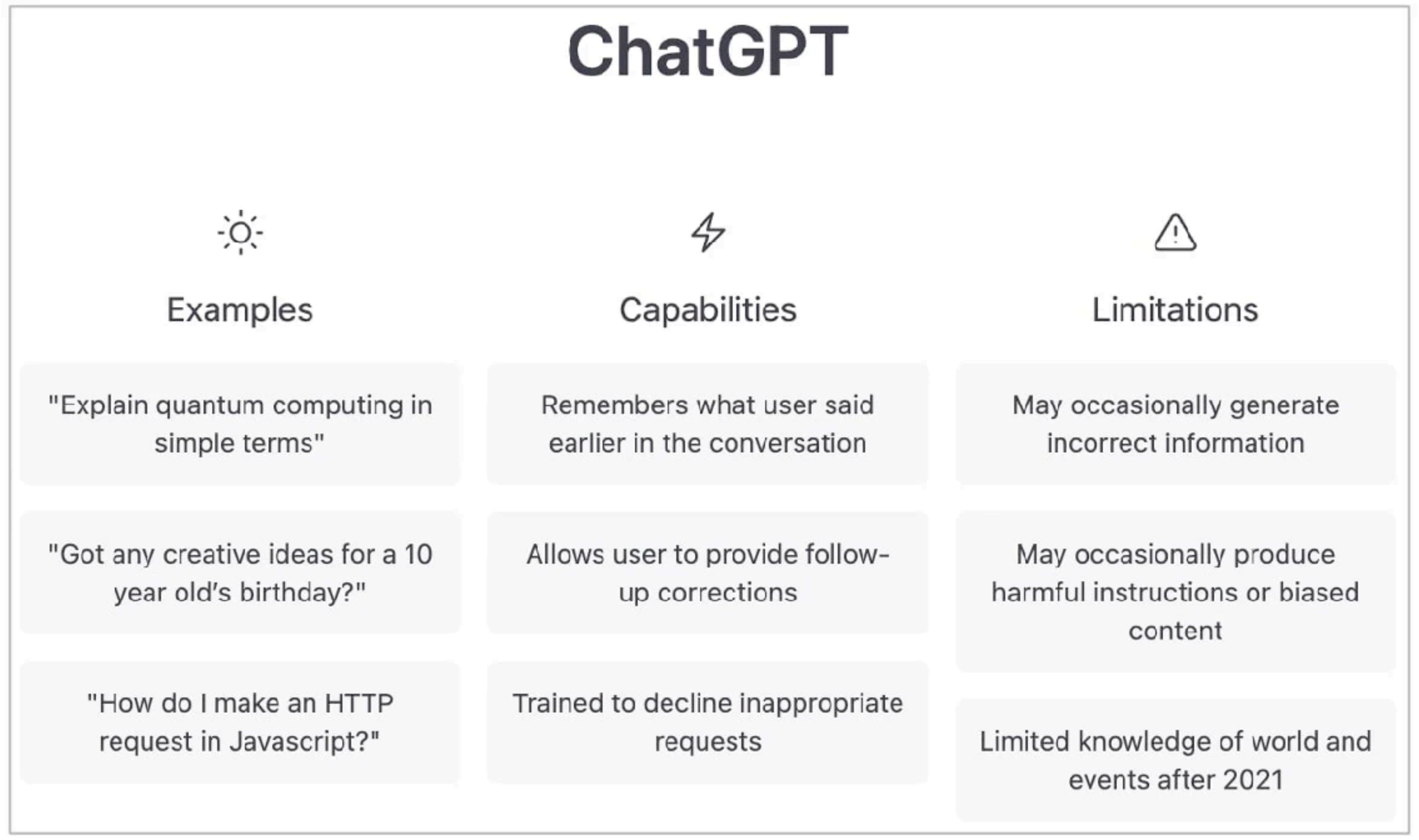
Hebrew University of Jerusalem

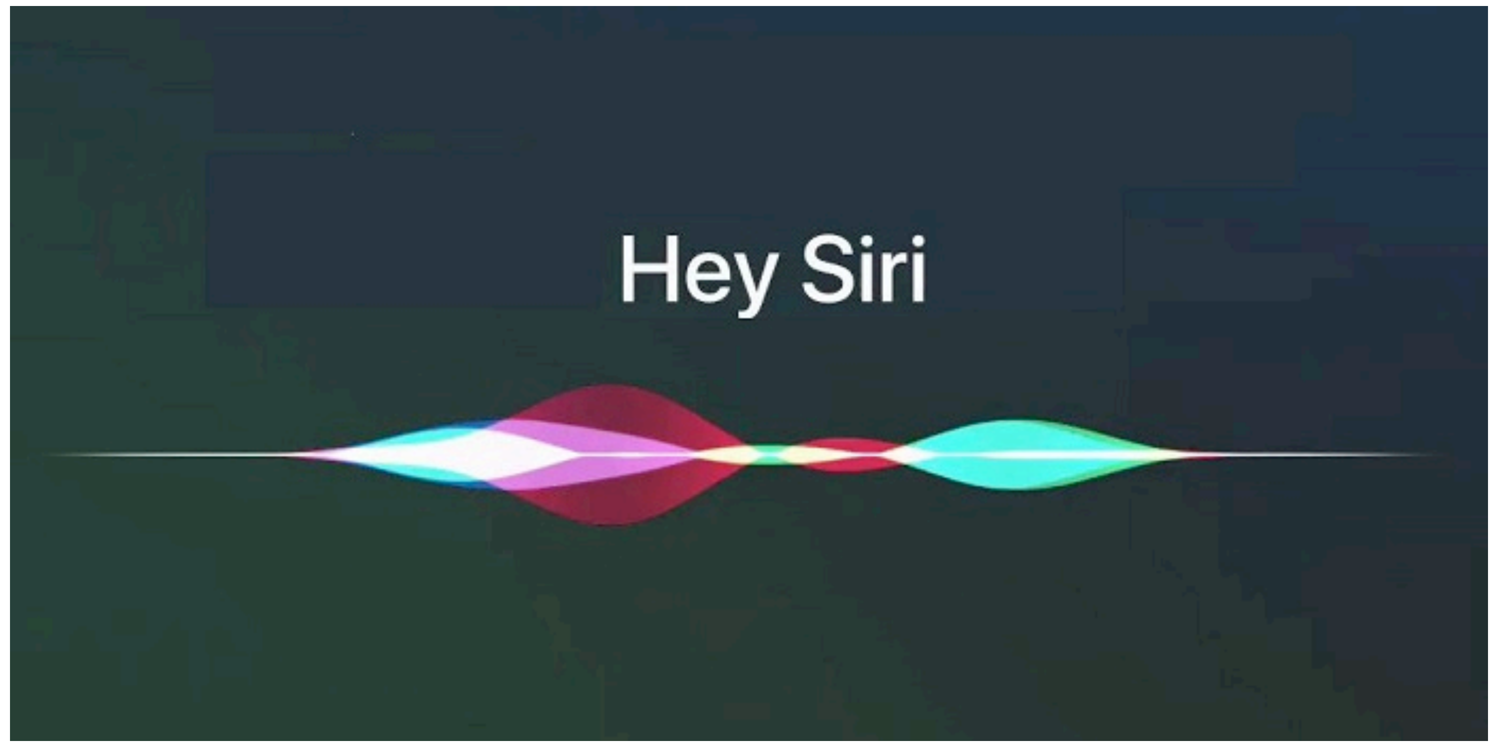
CS Colloquium, EPFL
December 2022

THE HEBREW UNIVERSITY OF JERUSALEM

AI Today

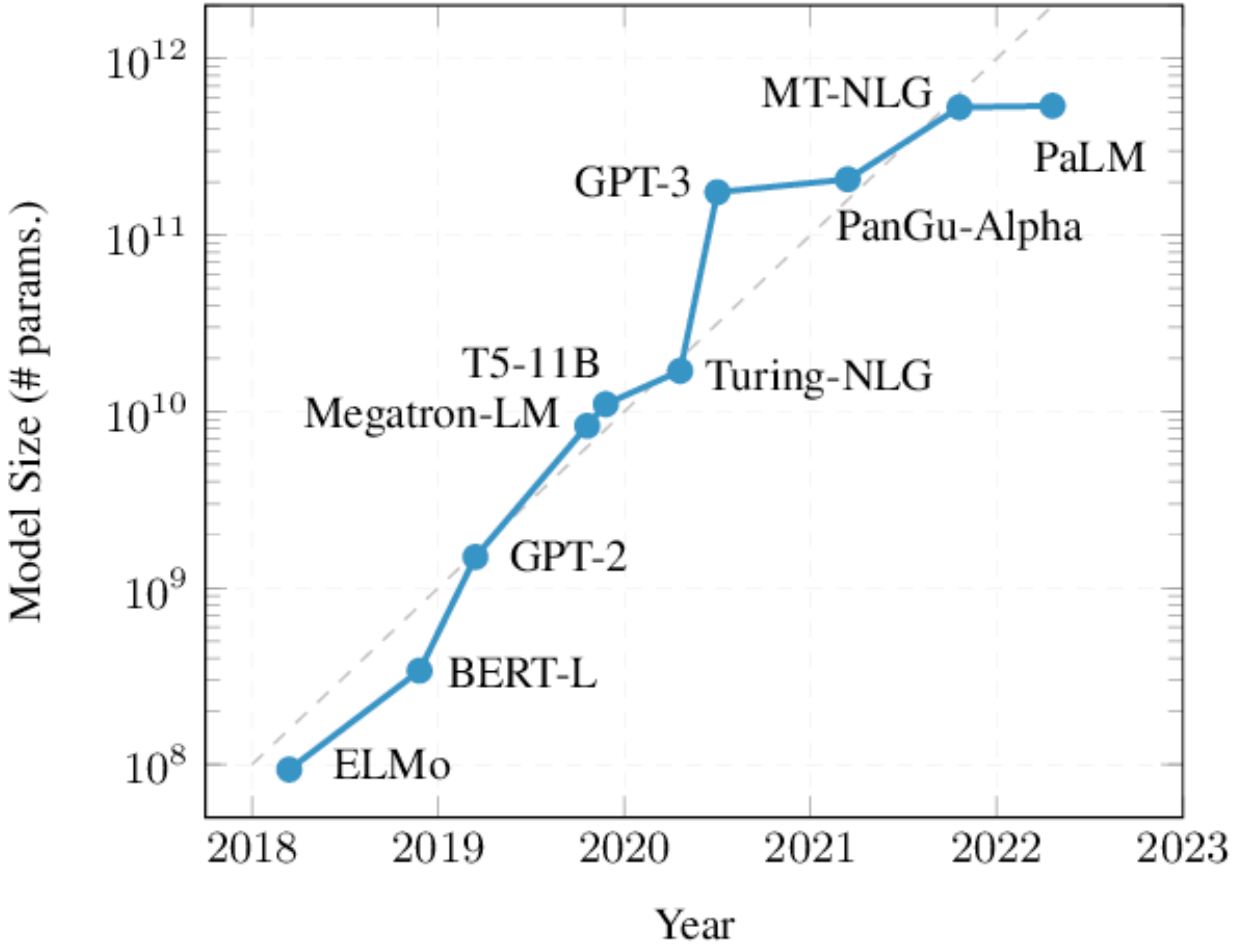
Microsoft Translator



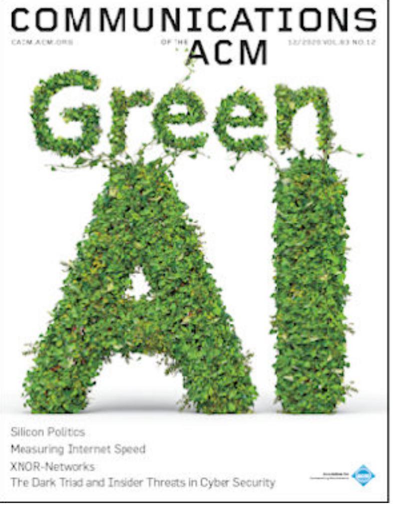


Scaling

5,000X in 4 Years



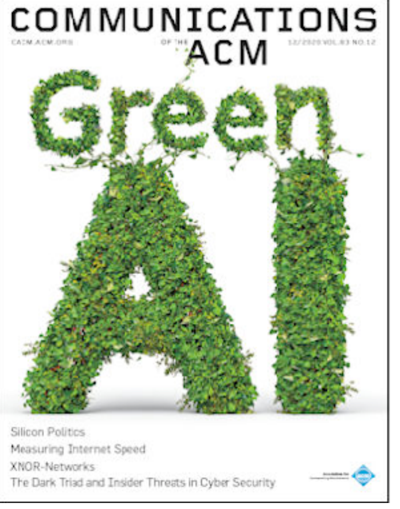
Taken from Lakim et al. (2022)



Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
    - Problems: inclusiveness, environment

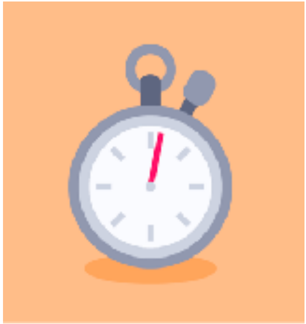


Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Problems: inclusiveness, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results



  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy

Problems with Scaling

Inclusiveness

Synced

AI TECHNOLOGY & INDUSTRY REVIEW

FEATURE ▾ INDUSTRY ▾ TECHNOLOGY COMMUNITY ▾ ABOUT US ▾ REPORT CONTRIBUTE TO SYNCED REVIEW

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** Cloud TPUs for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** V100 GPUs for approximately **1** day
- PaLM (Chowdhery et al., 2022) was trained on **6144** TPU v4 chips for **50** days and **3072** TPU v4 chips for **15** days

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** Cloud TPUs for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** V100 GPUs for approximately **1** day
- PaLM (Chowdhery et al., 2022) was trained on **6144** TPU v4 chips for **50** days and **3072** TPU v4 chips for **15** days

We need better reporting!

Number of Authors

Number of Authors

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Number of Authors

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

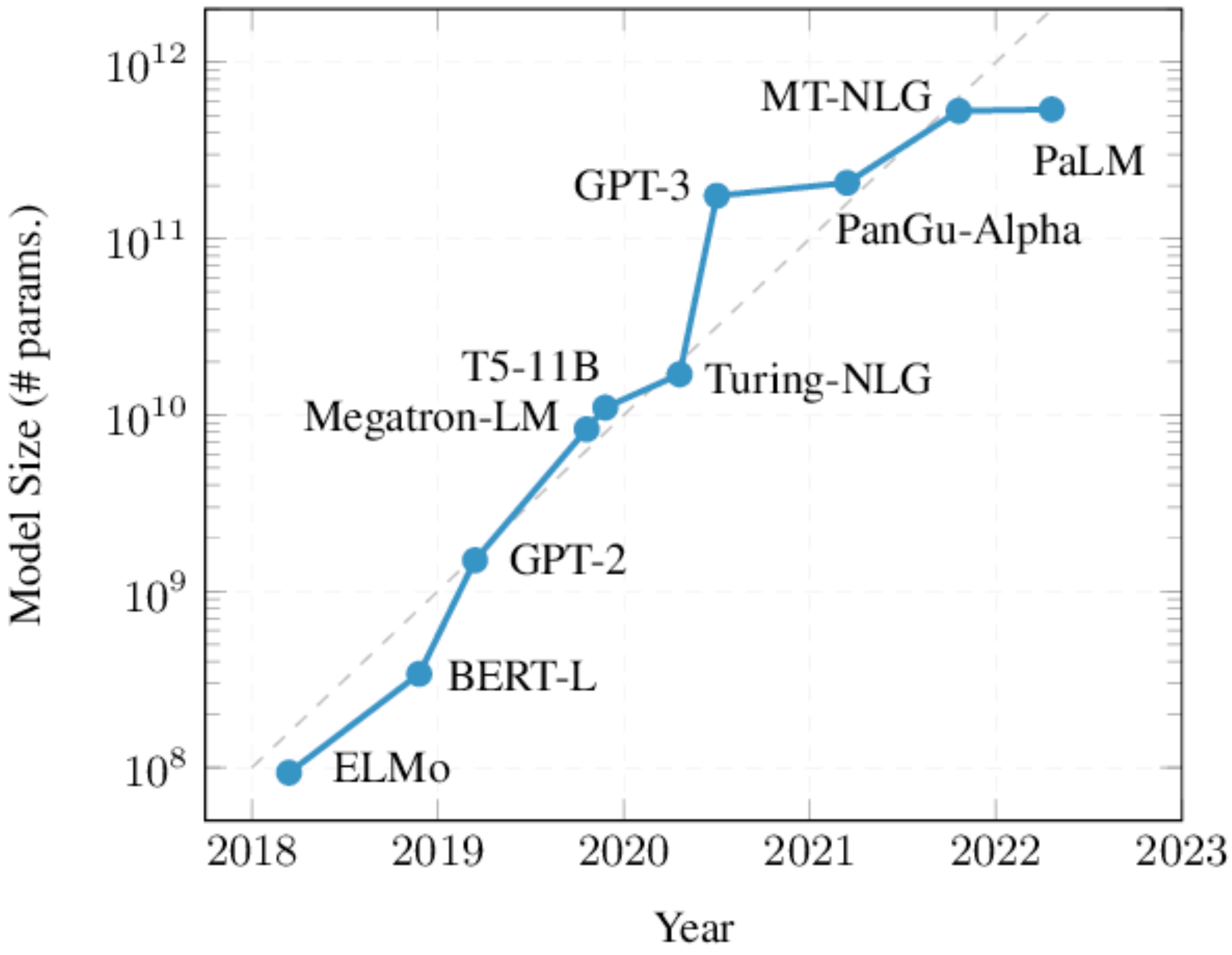
OpenAI

PaLM: Scaling Language Modeling with Pathways

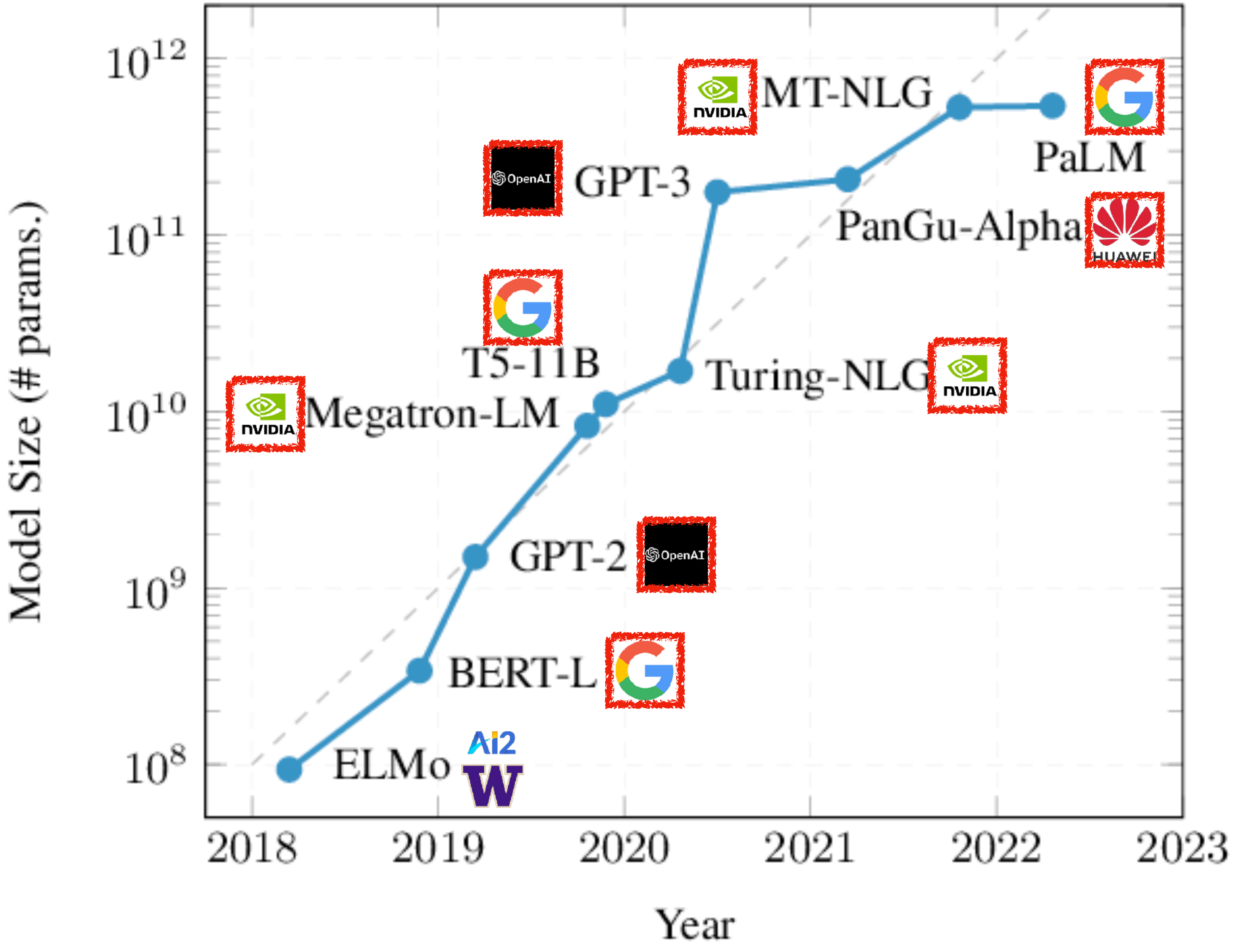
Aakanksha Chowdhery[*] Sharan Narang[*] Jacob Devlin[*]
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi
Sasha Tsvyashchenko Joshua Maynez Abhishek Rao[†] Parker Barnes Yi Tay
Noam Shazeer[‡] Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson
Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari
Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zhou Daphne Ippolito David Luan[‡] Hyeontaek Lim Barret Zoph
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz
Erica Moreira Rewon Child Oleksandr Polozov[†] Katherine Lee Zongwei Zhou
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta[†] Jason Wei
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

It's a Rich Man's World



It's a Rich Man's World



Problems with Scaling

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Problems with Scaling

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Is AI really creating an environmental problem?

Google's Answer: No!

BLOG ›

Good News About the Carbon Footprint of Machine Learning Training

TUESDAY, FEBRUARY 15, 2022

Posted by David Patterson, Distinguished Engineer, Google Research, Brain Team

Strubell et al.'s energy estimate for NAS ended up ***18.7X*** *too high for the average organization (…) and* ***88X*** *off in emissions for energy-efficient organizations like Google*

Google's Answer: No!


BLOG ›

Good News About the Carbon Footprint of Machine Learning Training

TUESDAY, FEBRUARY 15, 2022

Posted by David Patterson, Distinguished Engineer, Google Research, Brain Team

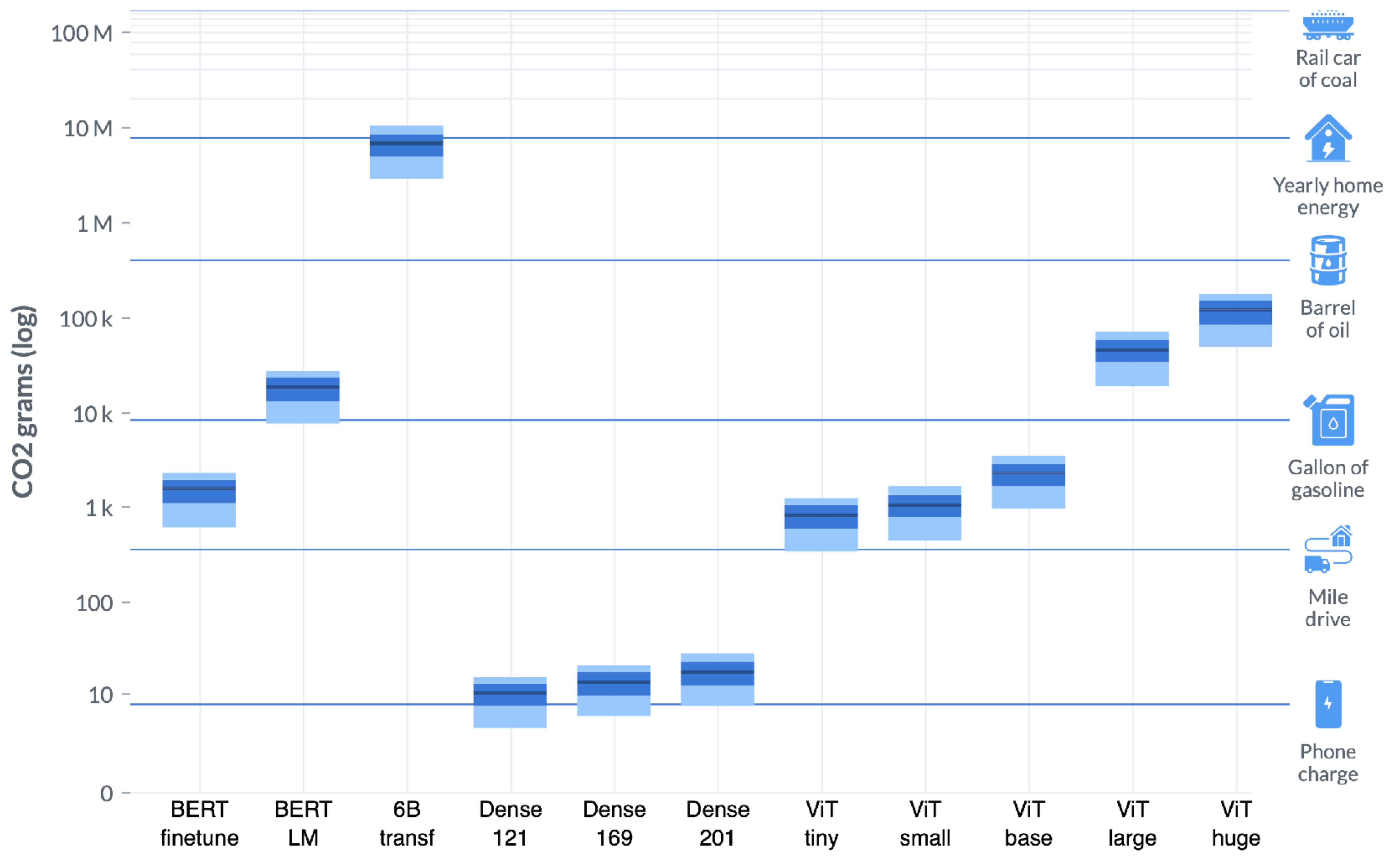
Strubell et al.'s energy estimate for NAS ended up ***18.7X*** *too high for the average organization (…) and* ***88X*** *off in emissions for energy-efficient organizations like Google*

We need better reporting!

Our Answer: Maybe?

Measuring the Carbon Intensity of AI in Cloud Instances

JESSE DODGE, Allen Institute for AI, USA
TAYLOR PREWITT, University of Washington, USA
REMI TACHET DES COMBES, Microsoft Research Montreal, USA
ERIKA ODMARK, Microsoft, USA
ROY SCHWARTZ, Hebrew University of Jerusalem, Israel
EMMA STRUBELL, Carnegie Mellon University, USA
ALEXANDRA SASHA LUCCIONI, Hugging Face, USA
NOAH A. SMITH, Allen Institute for AI and University of Washington, USA
NICOLE DECARIO, Allen Institute for AI, USA
WILL BUCHANAN, Microsoft, USA



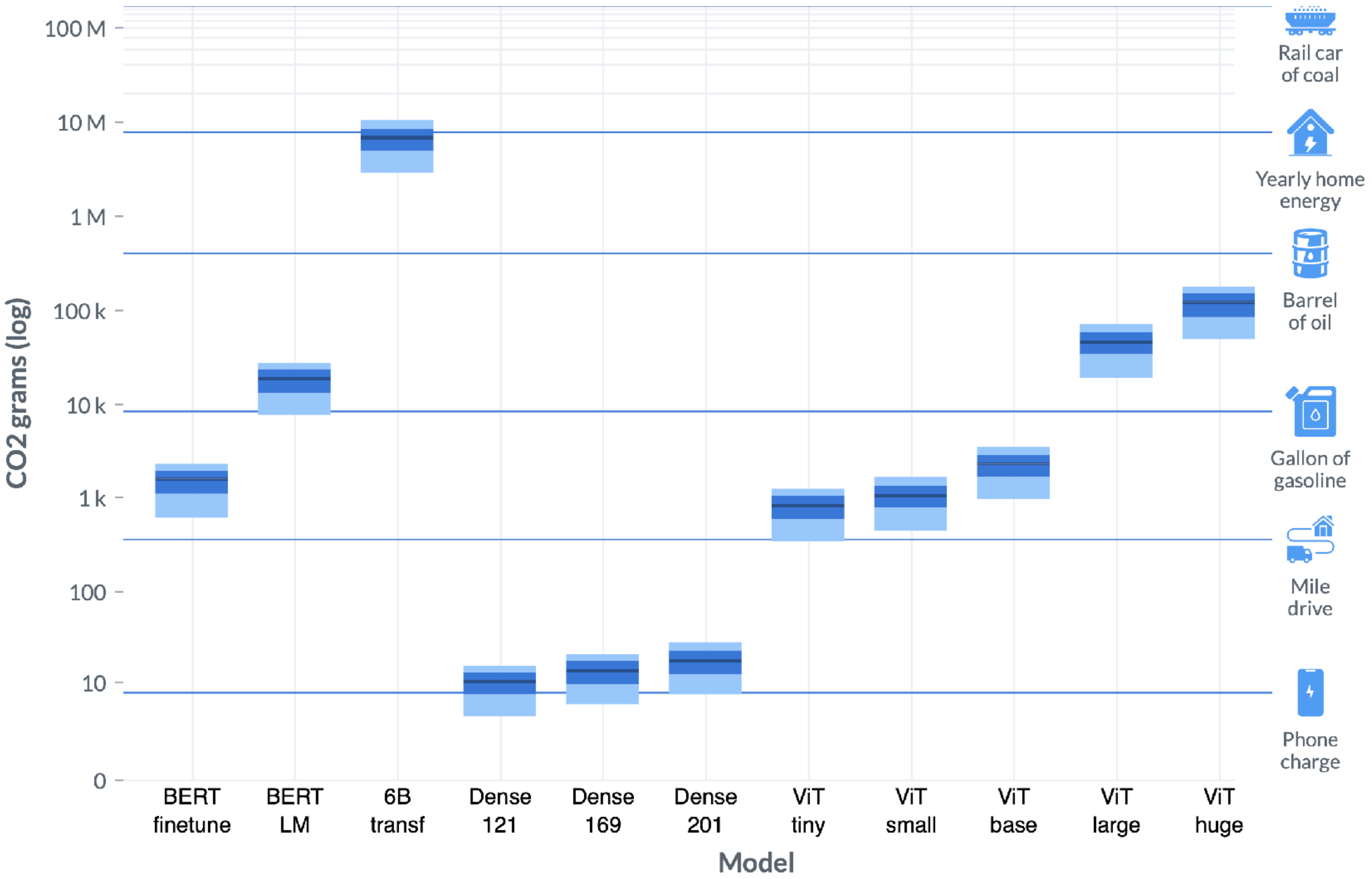
CO2 Relative Size Comparison

Green Software Foundation

Microsoft Azure

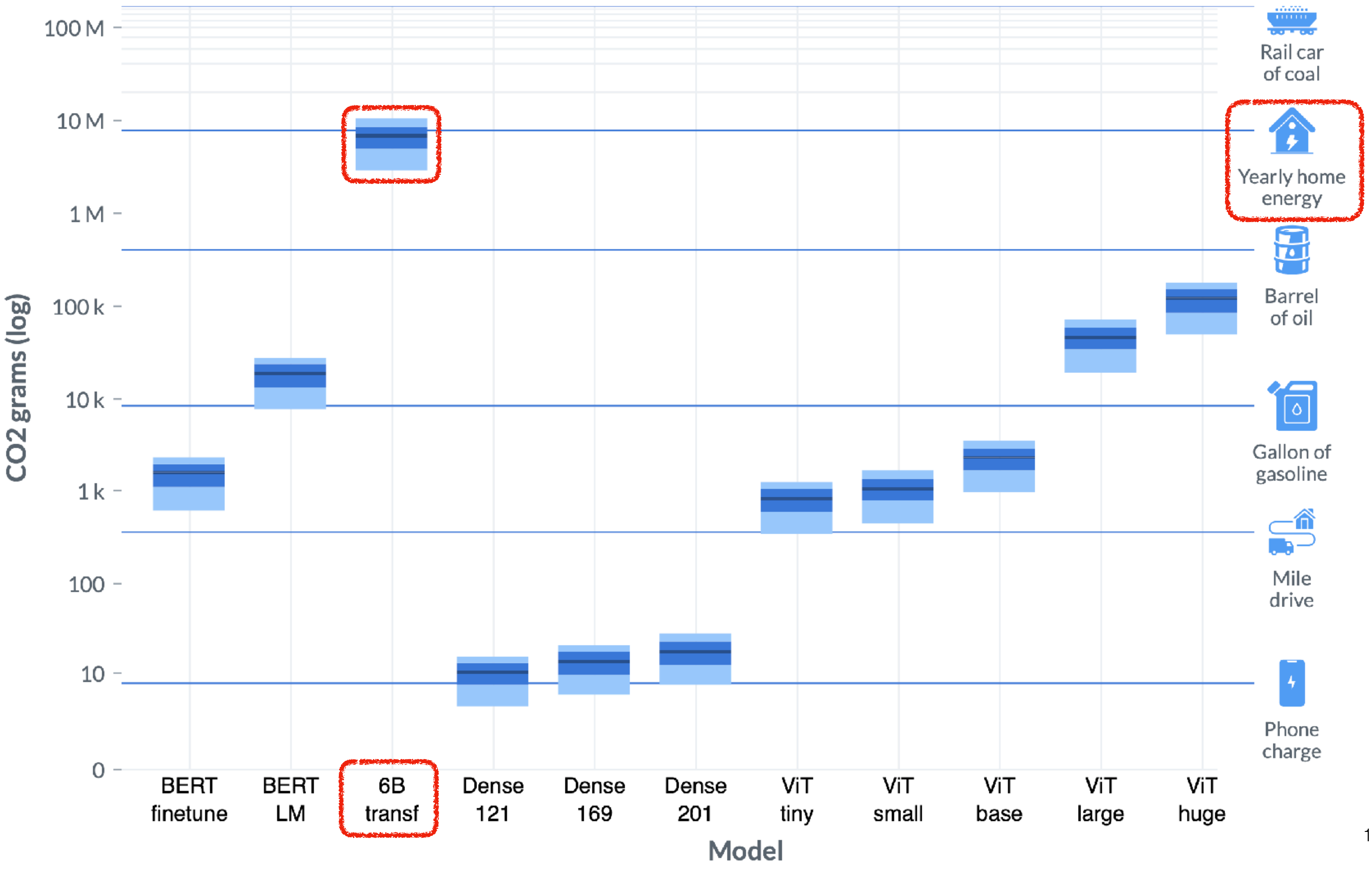
Measu

JESSE I
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



Measu

JESSE I
TAYLO
REMI T
ERIKA
ROY SC
EMMA
ALEXA
NOAH
NICOL
WILL B



AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - But these are typically run very few times

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - But these are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?

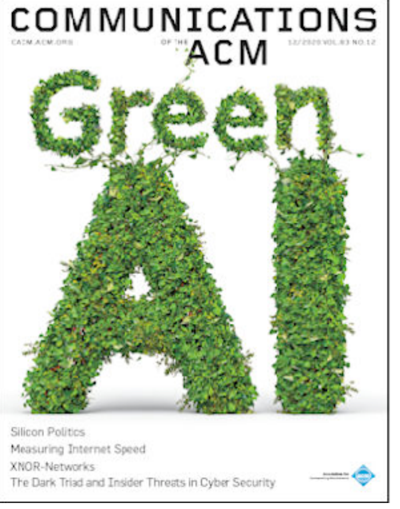
AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - But these are typically run very few times
- What about "normal" experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about **inference** operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - But these are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about **inference** operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

We need better reporting!

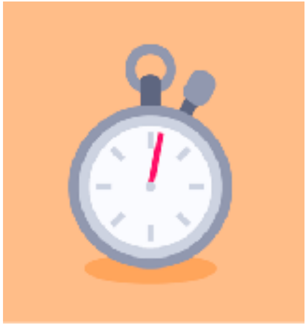


Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

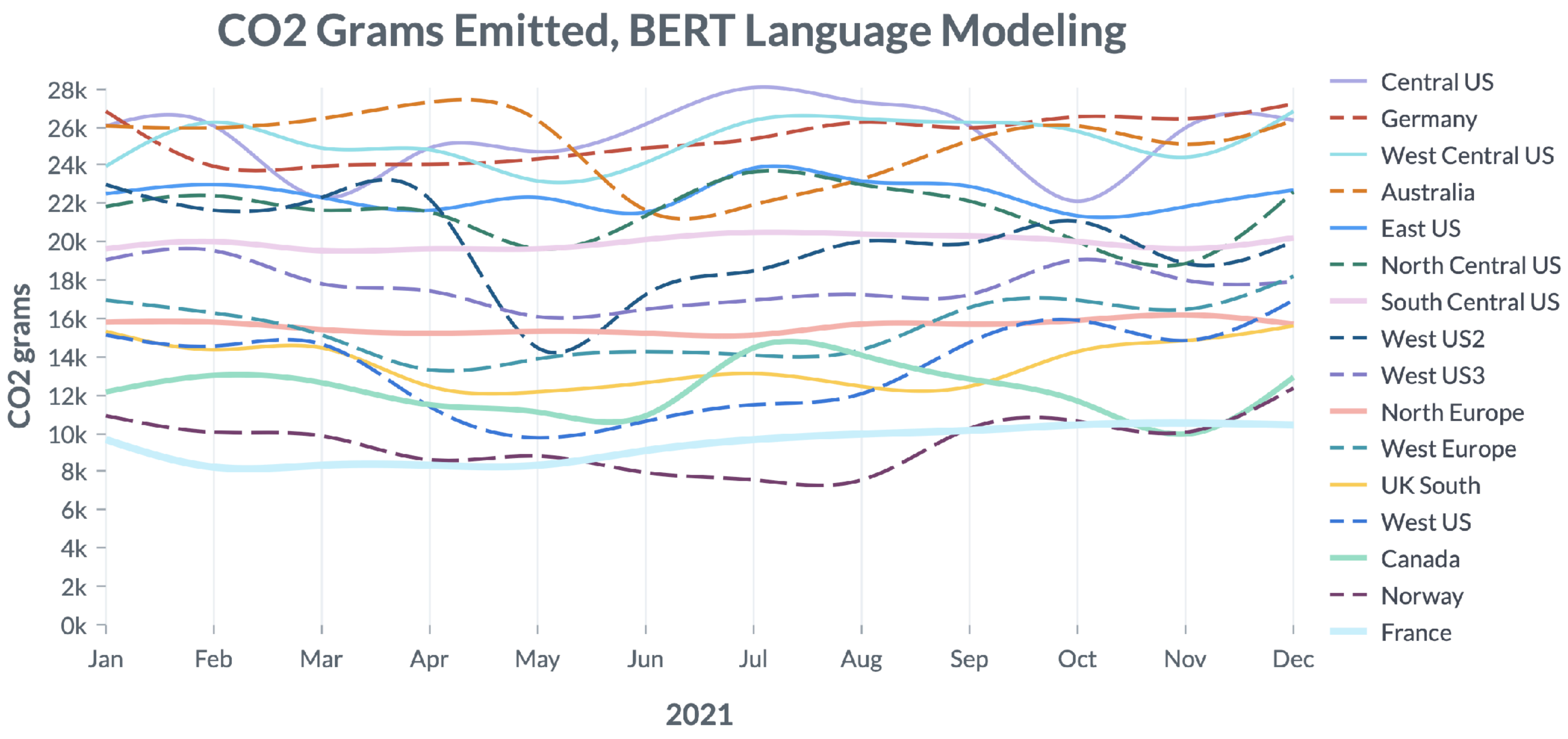
- Red AI
  - Problems: inclusiveness, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results



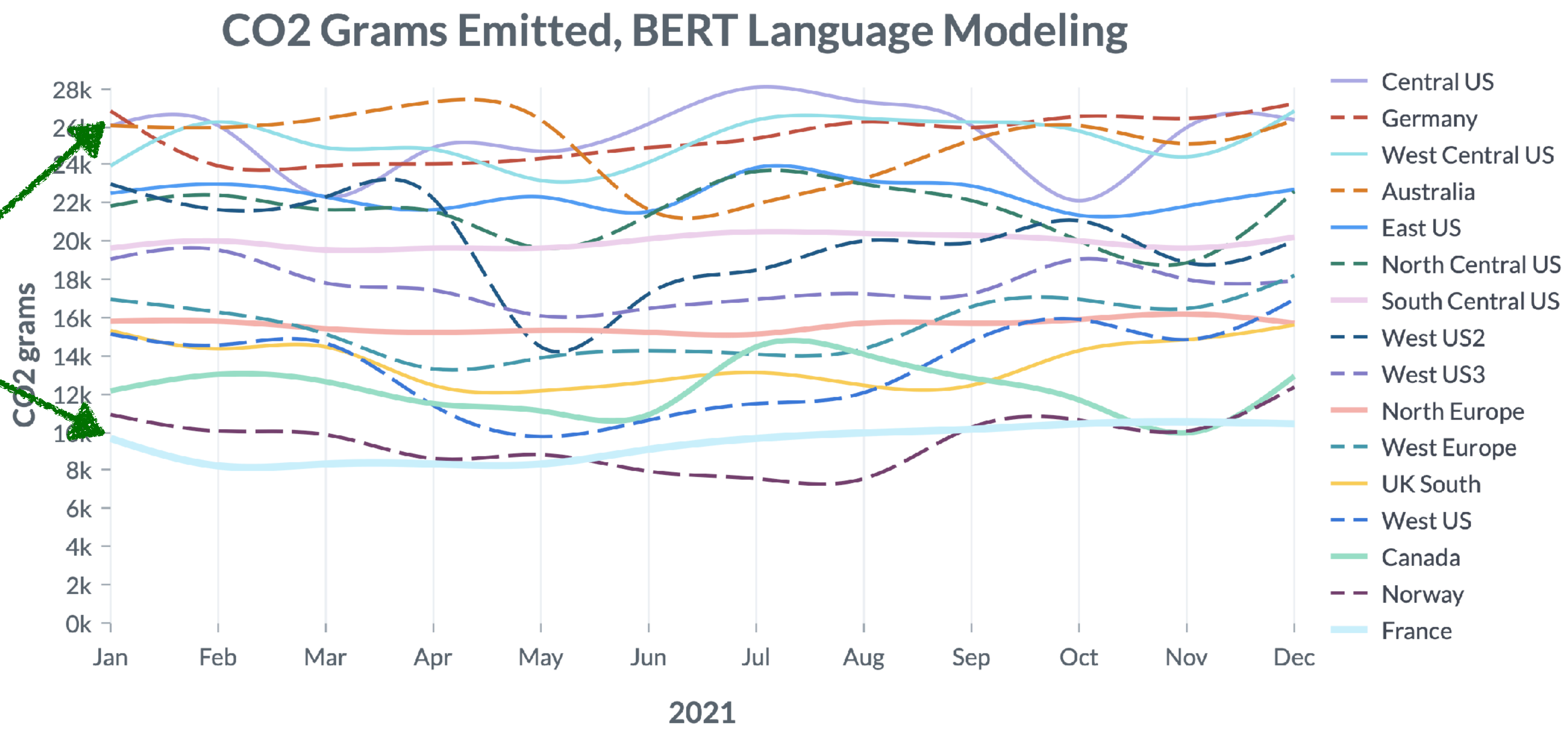
  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy

Cloud Location Matters



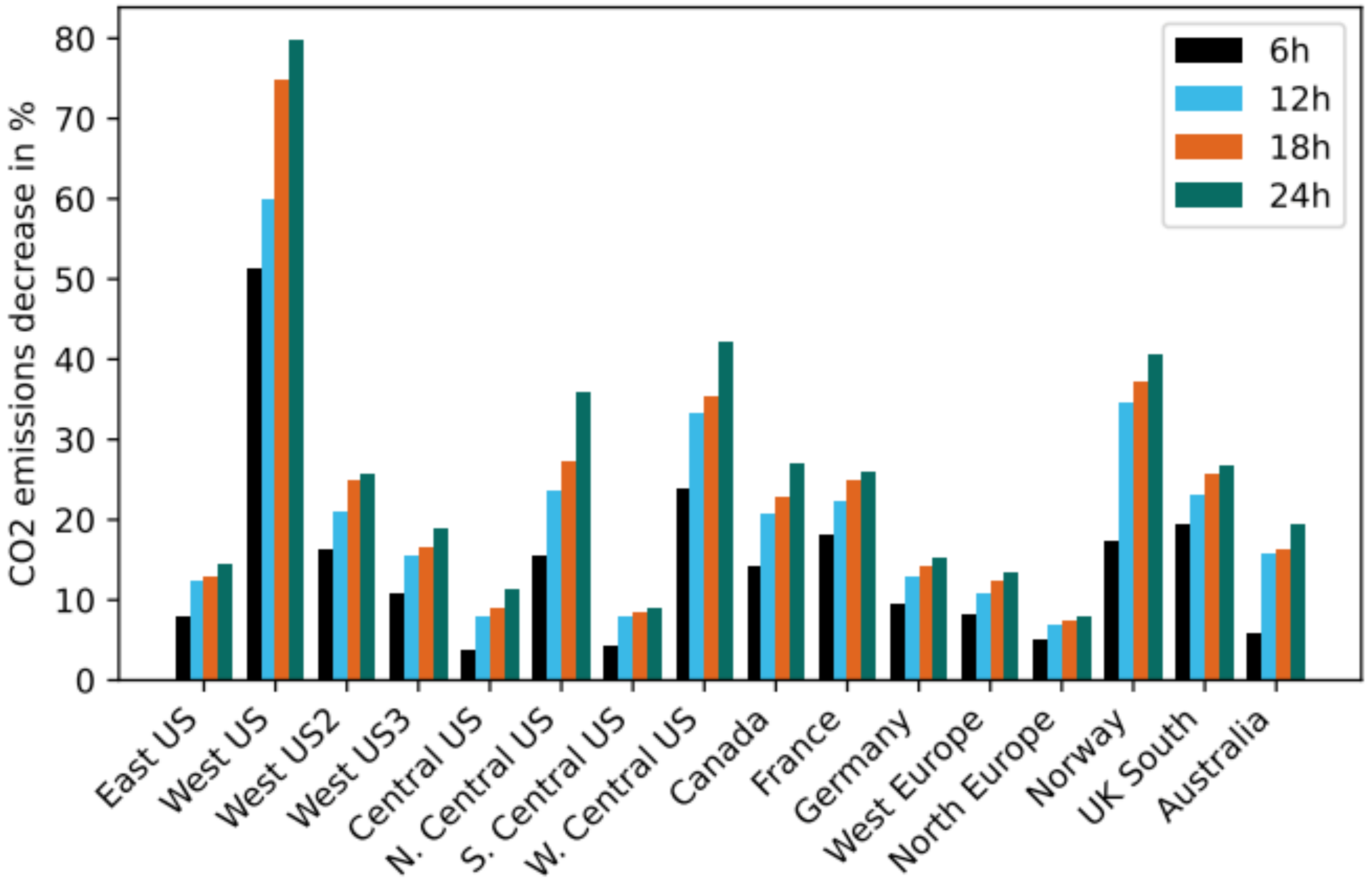
Cloud Location Matters

~3x difference

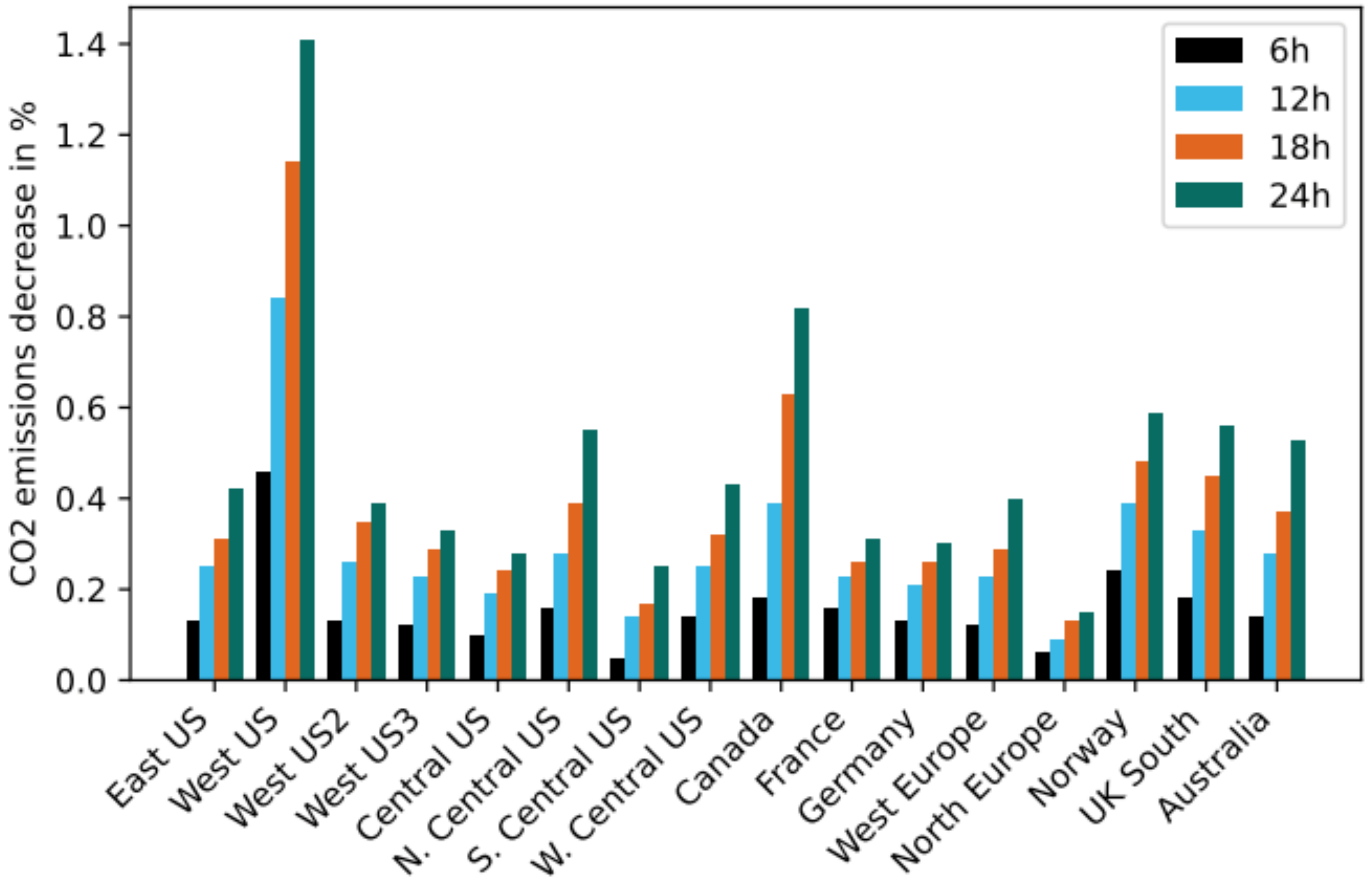


Time of Day Matters

Potential Saving with *Flexible Start*



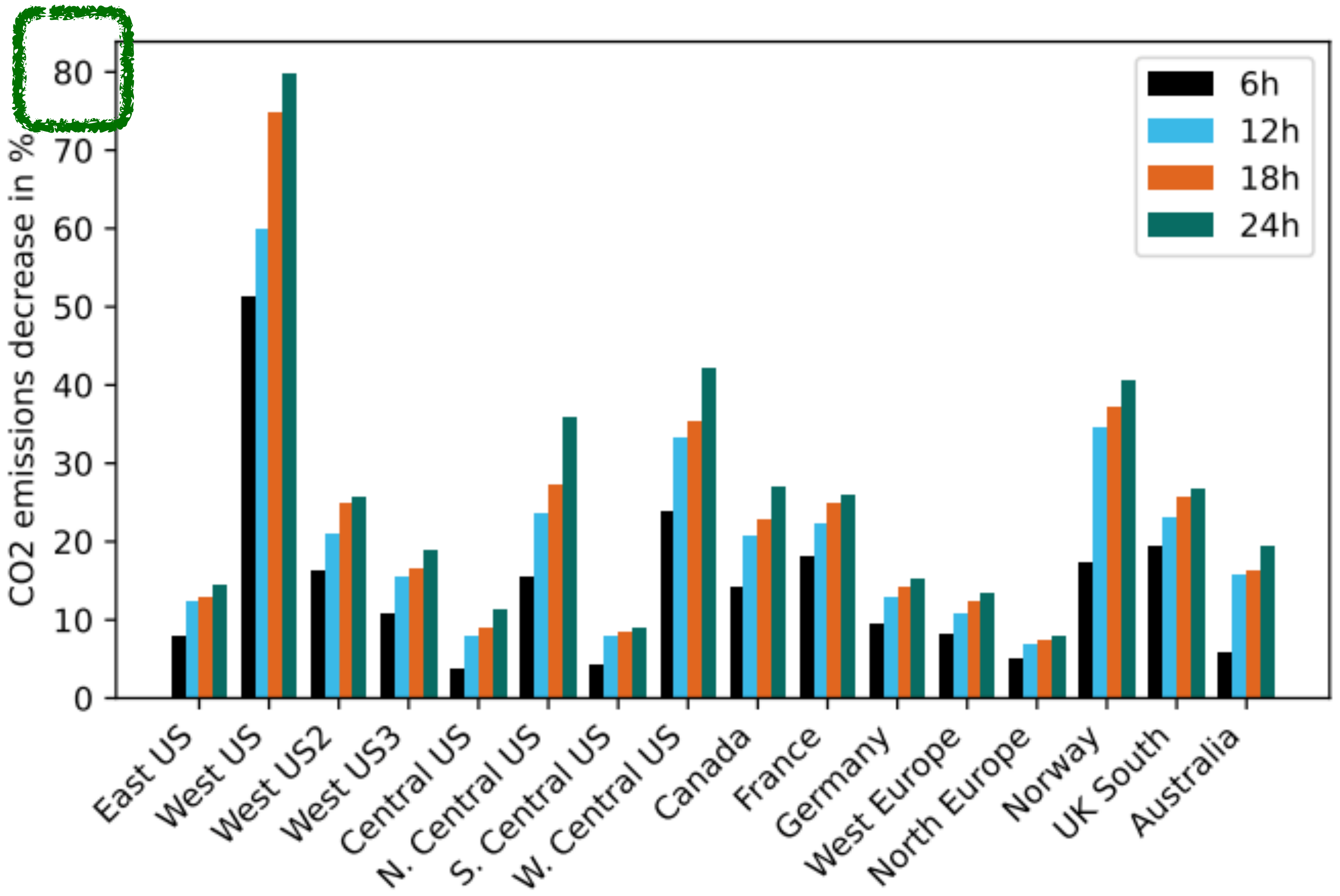
(a) *Flexible Start* optimization for Dense 201.



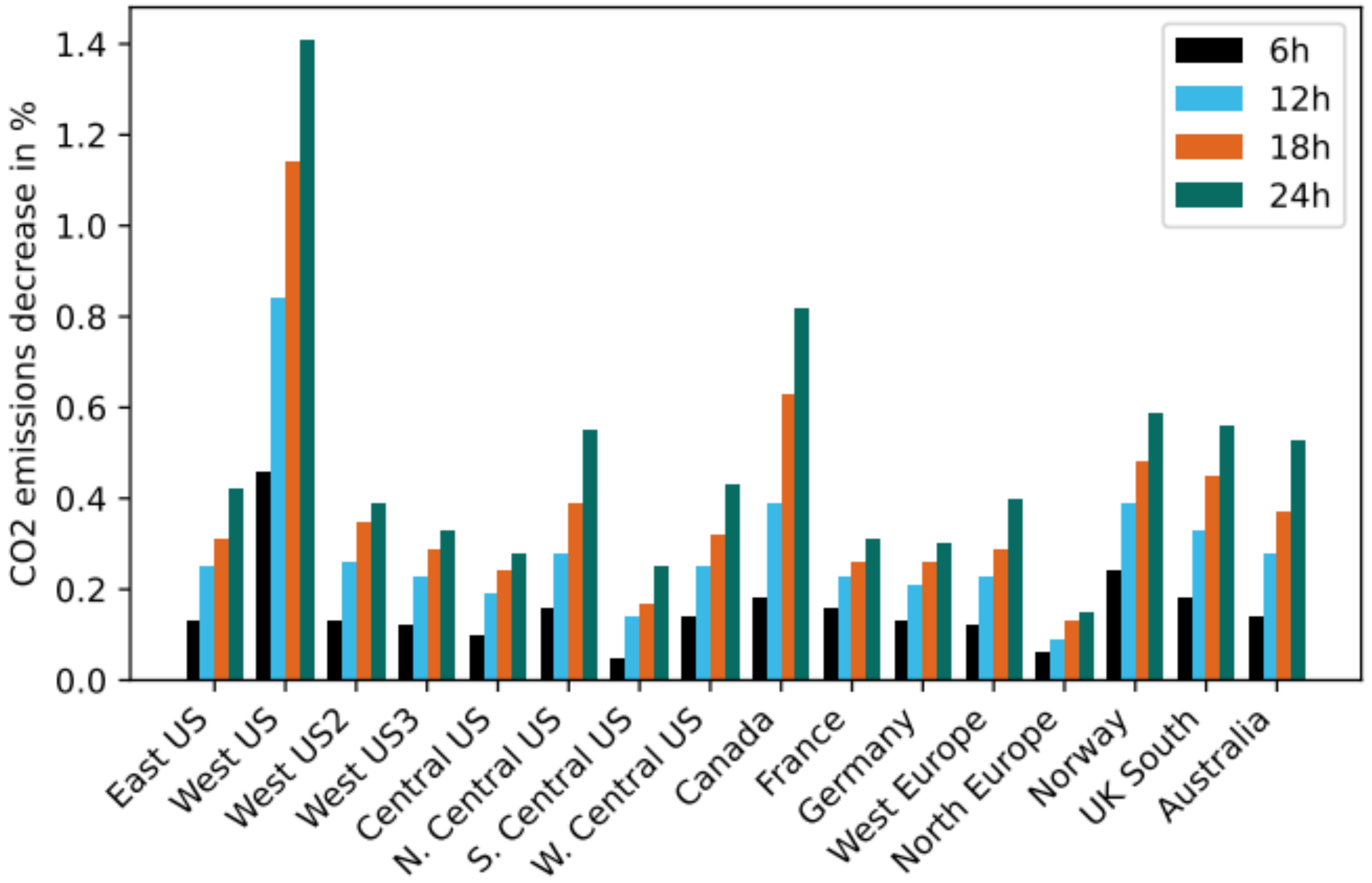
(b) *Flexible Start* optimization for 6B parameters Transformer.

Time of Day Matters

Potential Saving with *Flexible Start*



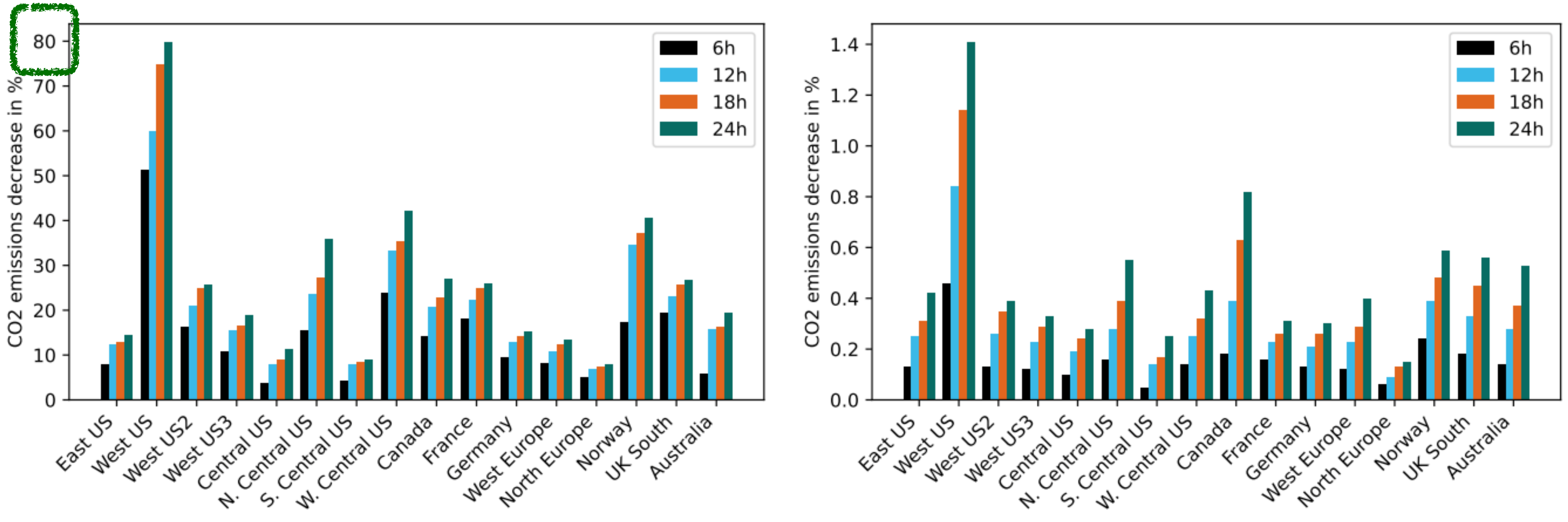
(a) *Flexible Start* optimization for Dense 201.



(b) *Flexible Start* optimization for 6B parameters Transformer.

Time of Day Matters

Potential Saving with *Flexible Start*



(a) *Flexible Start* optimization for Dense 201.

(b) *Flexible Start* optimization for 6B parameters Transformer.

We need better reporting!

Stop training large models?

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)

Large Models are Important

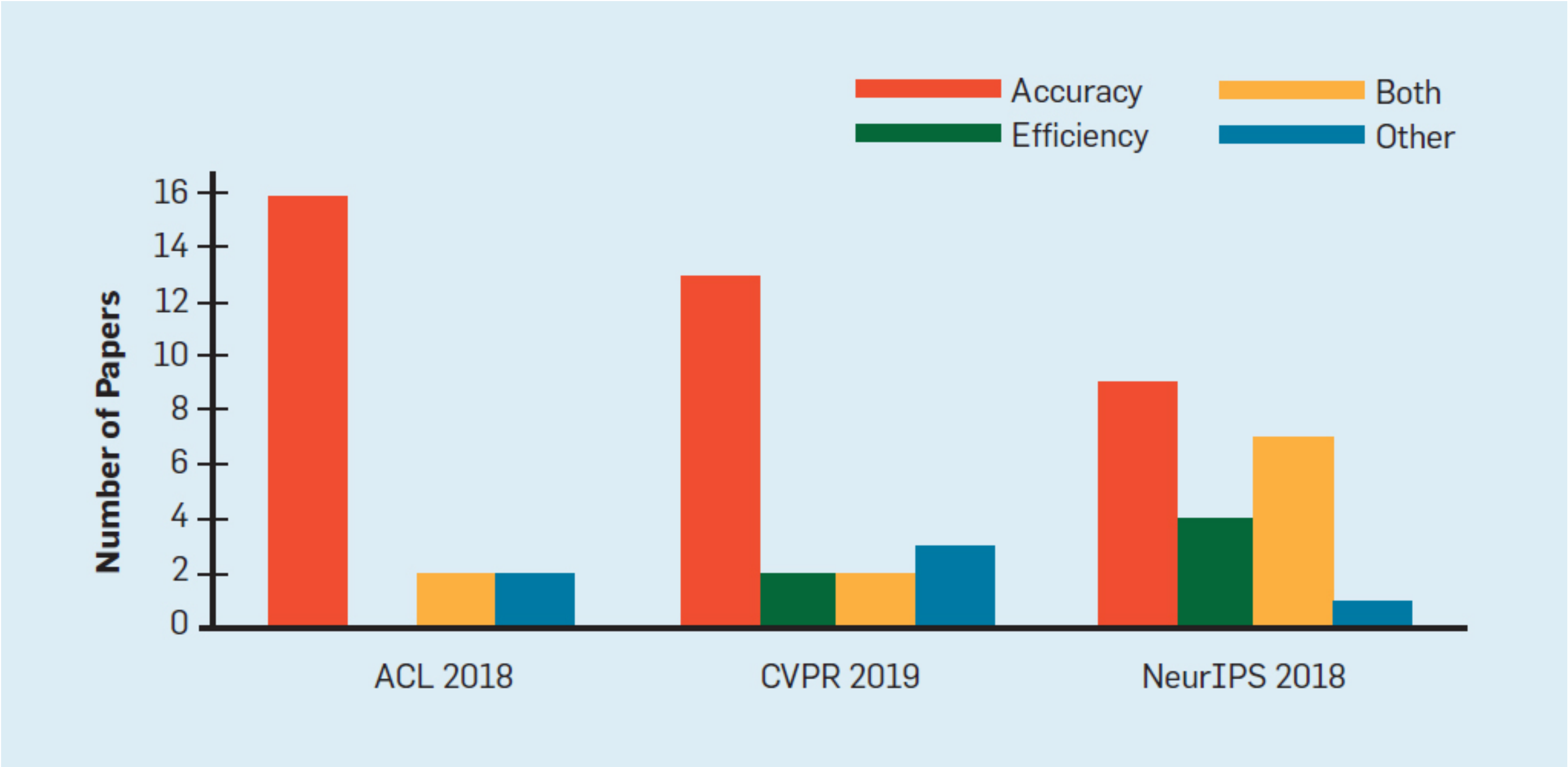
- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)
- But, large models have **concerning side affects**
  - Inclusiveness, environment

Large Models are Important

- Push the limits of SOTA
- Released large pre-trained models save compute
- Large models are potentially faster to train
  - Li et al. (2020)
- But, large models have **concerning side affects**
  - Inclusiveness, environment
- Our goal is to **mitigate these side affects**

Accuracy or Efficiency?

Accuracy or Efficiency?



S. et al. (2020)

Association for
Computational Linguistics

March 2022

Efficient NLP

Yuki Arase, Osaka University, arase@ist.osaka-u.ac.jp
Phil Blunsom, Oxford University, phil.blunsom@cs.ox.ac.uk
Mona Diab, George Washington University & Facebook AI, mtdiab@gwu.edu
Jesse Dodge, Allen Institute for AI, jessed@allenai.org
Iryna Gurevych, Technical University of Darmstadt, gurevych@ukp.informatik.tu-darmstadt.de
Percy Liang, Stanford University, pliang@cs.stanford.edu
Colin Raffel, UNC & HuggingFace, craffel@gmail.com
Andreas Rücklé, Amazon, andreas@rueckle.net
Roy Schwartz, The Hebrew University of Jerusalem, roy.schwartz1@mail.huji.ac.il
Noah A. Smith, University of Washington & Allen Institute for AI, nasmith@cs.washington.edu
Emma Strubell, Carnegie Mellon University & Google, strubell@cmu.edu
Yue Zhang, Westlake University, yue.zhang@wias.org.cn

Association for
Computational Linguistics

March 2022

Efficient NLP

Yuki Arase, Osaka University, arase@ist.osaka-u.ac.jp
Phil Blunsom, Oxford University, phil.blunsom@cs.ox.ac.uk
Mona Diab, George Washington University & Facebook AI, mtdiab@gwu.edu
Jesse Dodge, Allen Institute for AI, jessed@allenai.org
Iryna Gurevych, Technical University of Darmstadt, gurevych@ukp.informatik.tu-darmstadt.de
Percy Liang, Stanford University, pliang@cs.stanford.edu
Colin Raffel, UNC & HuggingFace, craffel@gmail.com
Andreas Rücklé, Amazon, andreas@rueckle.net
Roy Schwartz, The Hebrew University of Jerusalem, roy.schwartz1@mail.huji.ac.il
Noah A. Smith, University of Washington & Allen Institute for AI, nasmith@cs.washington.edu
Emma Strubell, Carnegie Mellon University & Google, strubell@cmu.edu
Yue Zhang, Westlake University, yue.zhang@wias.org.cn

- Setting up conference areas that target efficiency

Association for
Computational Linguistics

March 2022

Efficient NLP

Yuki Arase, Osaka University, arase@ist.osaka-u.ac.jp
Phil Blunsom, Oxford University, phil.blunsom@cs.ox.ac.uk
Mona Diab, George Washington University & Facebook AI, mtdiab@gwu.edu
Jesse Dodge, Allen Institute for AI, jessed@allenai.org
Iryna Gurevych, Technical University of Darmstadt, gurevych@ukp.informatik.tu-darmstadt.de
Percy Liang, Stanford University, pliang@cs.stanford.edu
Colin Raffel, UNC & HuggingFace, craffel@gmail.com
Andreas Rücklé, Amazon, andreas@rueckle.net
Roy Schwartz, The Hebrew University of Jerusalem, roy.schwartz1@mail.huji.ac.il
Noah A. Smith, University of Washington & Allen Institute for AI, nasmith@cs.washington.edu
Emma Strubell, Carnegie Mellon University & Google, strubell@cmu.edu
Yue Zhang, Westlake University, yue.zhang@wias.org.cn

- Setting up conference areas that target efficiency
- Encouraging the release of trained models

About Dagstuhl | **Program** | Publications

You are here: **Program** » **Seminar Calendar** » Seminar Homepage

https://www.dagstuhl.de/22232

June 6 – 10 , 2022, Dagstuhl Seminar 22232

Efficient and Equitable Natural Language Processing in the Age of Deep Learning

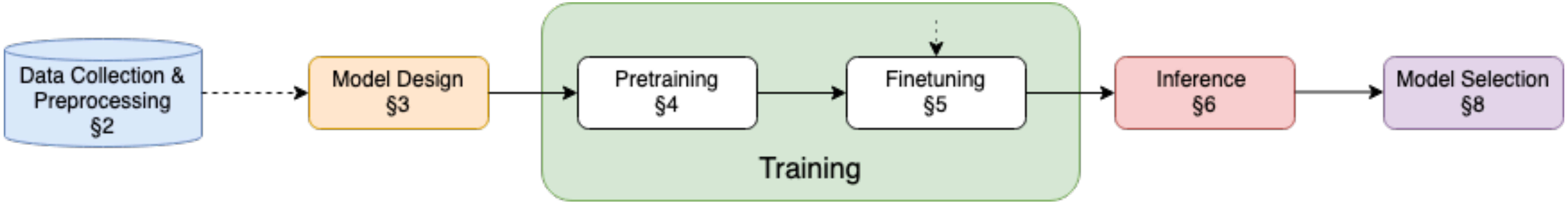
Organizers

Jesse Dodge (AI2 – Seattle, US)
Iryna Gurevych (TU Darmstadt, DE)
Roy Schwartz (The Hebrew University of Jerusalem, IL)
Emma Strubell (Carnegie Mellon University – Pittsburgh, US)

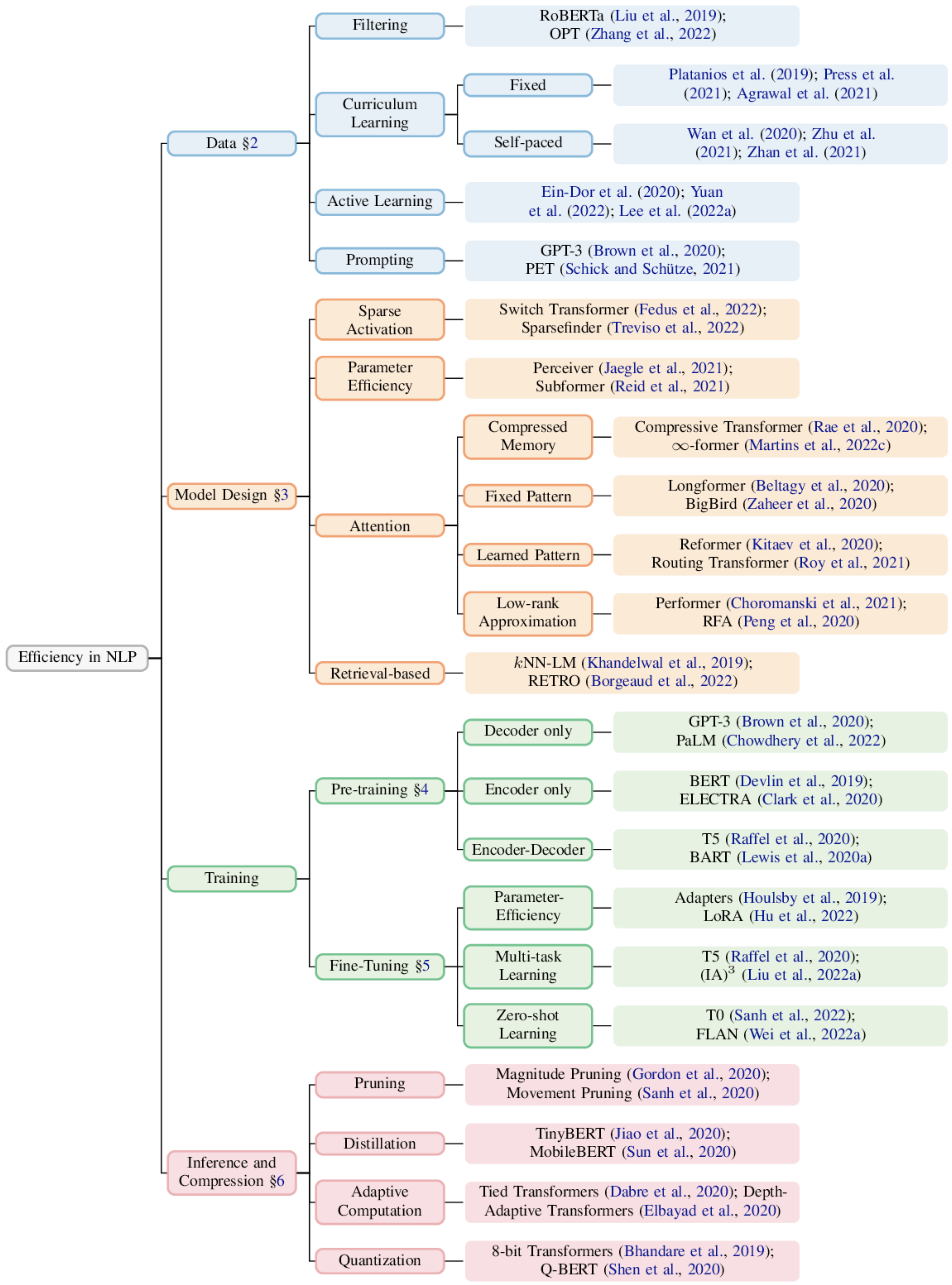
Efficient Methods for Natural Language Processing: A Survey

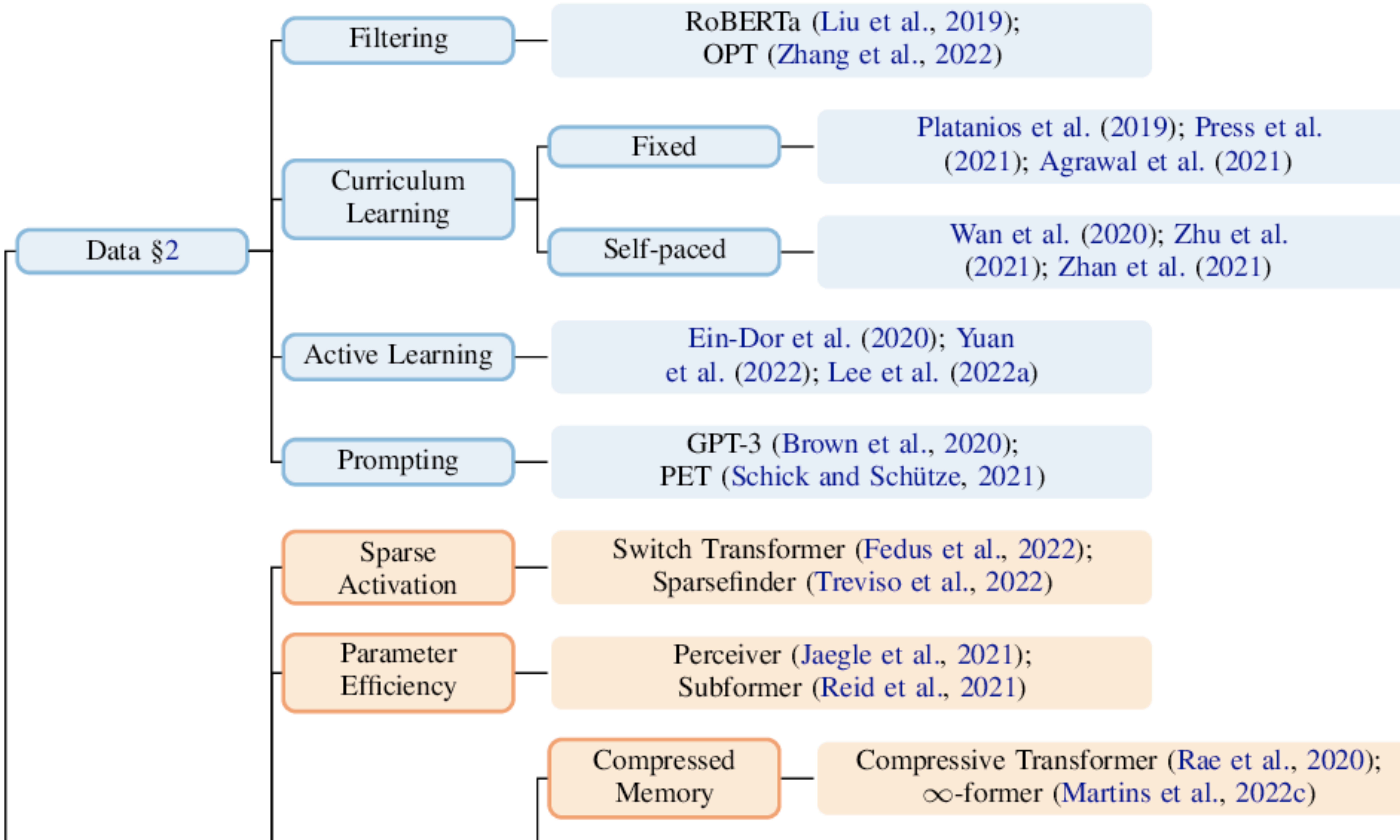
Marcos Treviso[10*], **Tianchu Ji**[3*], **Ji-Ung Lee**[7*], **Betty van Aken**[8], **Qingqing Cao**[2],
Manuel R. Ciosici[9], **Michael Hassid**[1], **Kenneth Heafield**[13], **Sara Hooker**[5],
Pedro H. Martins[10], **André F. T. Martins**[10], **Peter Milder**[3], **Colin Raffel**[6],
Edwin Simpson[4], **Noam Slonim**[12], **Niranjan Balasubramanian**[3], **Leon Derczynski**[11], **Roy Schwartz**[1]
[1]The Hebrew University of Jerusalem, [2]University of Washington, [3]Stony Brook University,
[4]University of Bristol, [5]Cohere For AI, [6]University of North Carolina at Chapel Hill,
[7]Technical University of Darmstadt, [8]Berliner Hochschule für Technik,
[9]University of Southern California, [10]IST/University of Lisbon & Instituto de Telecomunicações,
[11]IT University of Copenhagen, [12]IBM Research, [13]University of Edinburgh

Efficient Methods in NLP



Efficient Methods in NLP





Data §2
Filtering
RoBERTa (Liu et al., 2019); OPT (Zhang et al., 2022)
Curriculum Learning
Fixed
Platanios et al. (2019); Press et al. (2021); Agrawal et al. (2021)
Self-paced
Wan et al. (2020); Zhu et al. (2021); Zhan et al. (2021)
Active Learning
Ein-Dor et al. (2020); Yuan et al. (2022); Lee et al. (2022a)
Prompting
GPT-3 (Brown et al., 2020); PET (Schick and Schütze, 2021)
Sparse Activation
Switch Transformer (Fedus et al., 2022); Sparsefinder (Treviso et al., 2022)
Parameter Efficiency
Perceiver (Jaegle et al., 2021); Subformer (Reid et al., 2021)
Compressed Memory
Compressive Transformer (Rae et al., 2020); ∞-former (Martins et al., 2022c)

Filtering

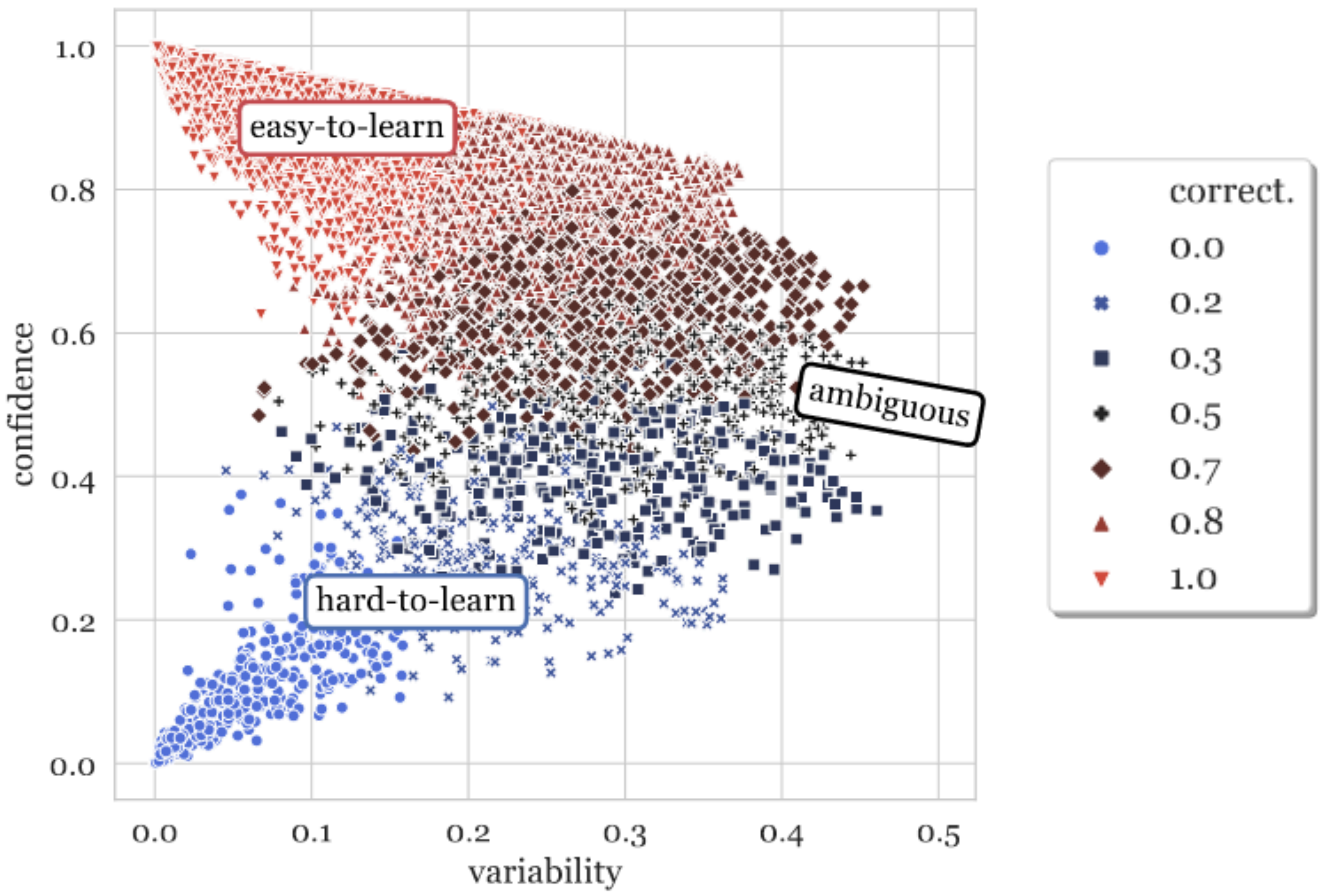
- Non-text
  - Gibrish, HTML
- Text in other languages
- Foul text
- Typically done via simple, rule-based heuristics
  - Noisy process

Smart Filtering

Swayamdipta, S. et al., EMNLP 2020

- Not all training instances contribute the same to learning
  - Some are "*easy-to-learn*", others are more challenging

Dataset Map



easy-to-learn
ambiguous
hard-to-learn
confidence
variability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
correct.
0.0
0.2
0.3
0.5
0.7
0.8
1.0

Experiments

WinoGrande, RoBERTa-Large

	WinoG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

	WinoG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

		WinoG. Val. (ID)	WSC (OOD)
	100% train	$79.7_{0.2}$	$86.0_{0.1}$
33%	random	$73.3_{1.3}$	$85.6_{0.4}$
	ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Experiments

WinoGrande, RoBERTa-Large

	WinoG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

33%

Experiments

WinoGrande, RoBERTa-Large

	WinoG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

33%

Experiments

WinoGrande, RoBERTa-Large

	WinoG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

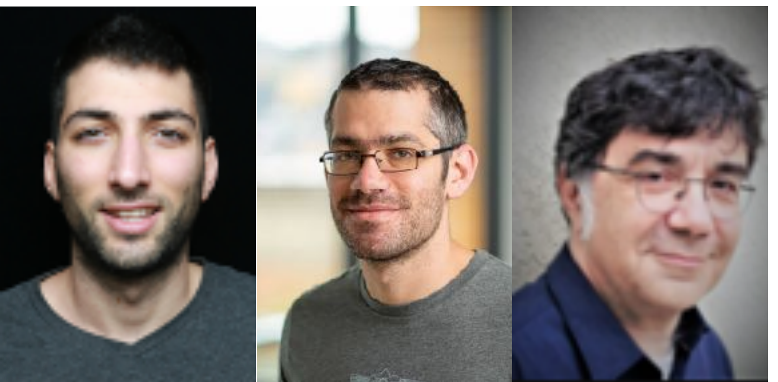
33%

Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

Baseline MLM

A tiger [MASK] eating the carrot



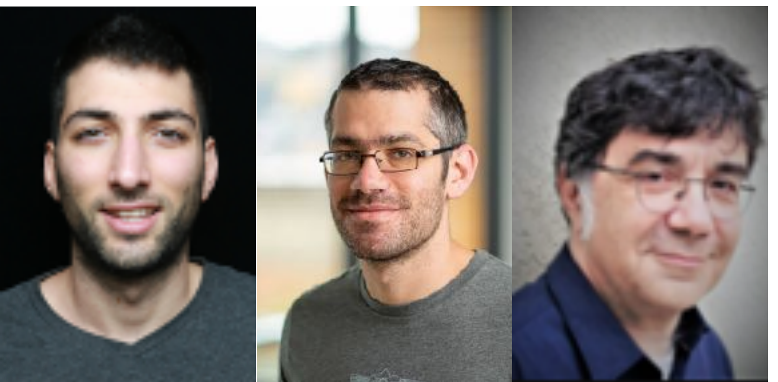
Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

- Current practice: **randomly** mask some of the words in the sentence
    - Many of them are **stop words** and **punctuation**

Baseline MLM

A tiger [MASK] eating the carrot



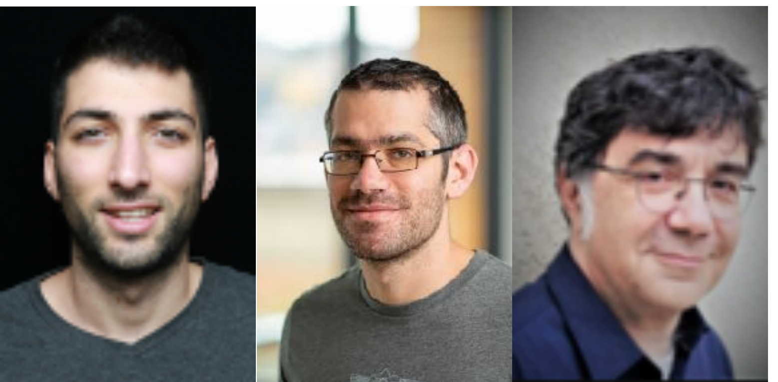
Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

- Current practice: **randomly** mask some of the words in the sentence
  - Many of them are **stop words** and **punctuation**
- Our proposal: only mask **content words**

Baseline MLM

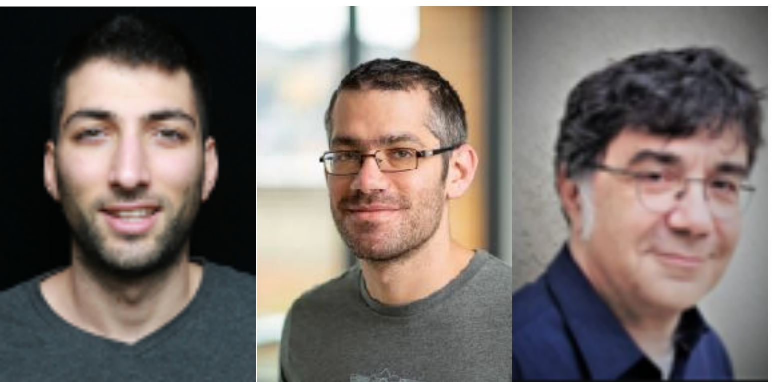
A tiger [MASK] eating the carrot



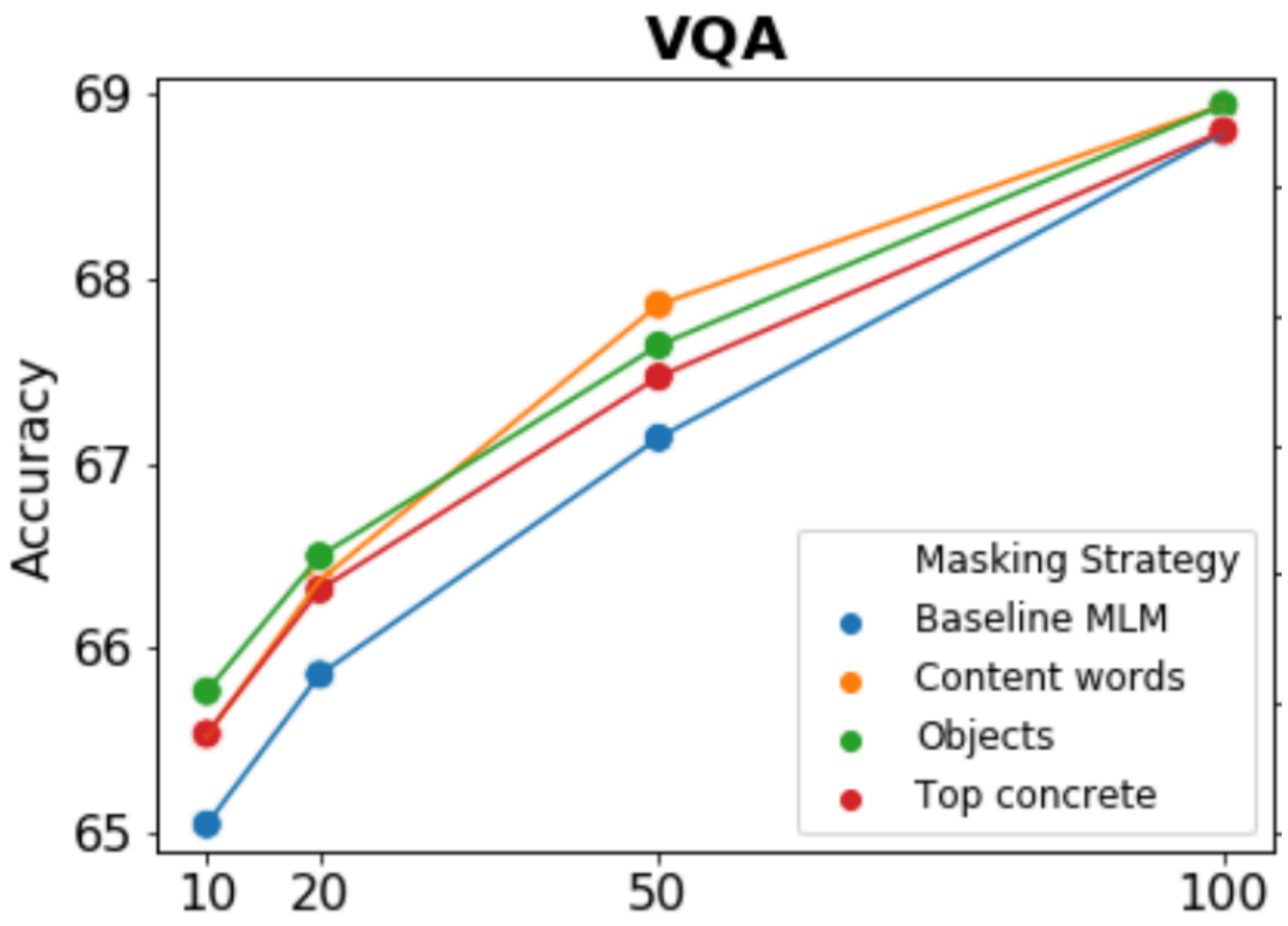
Data Efficient Masked Language Modeling for Vision and Language

Bitton, Stanovsky, Elhadad & S., Findings of EMNLP 2021

- Current practice: **randomly** mask some of the words in the sentence
    - Many of them are **stop words** and **punctuation**
- Our proposal: only mask **content words**

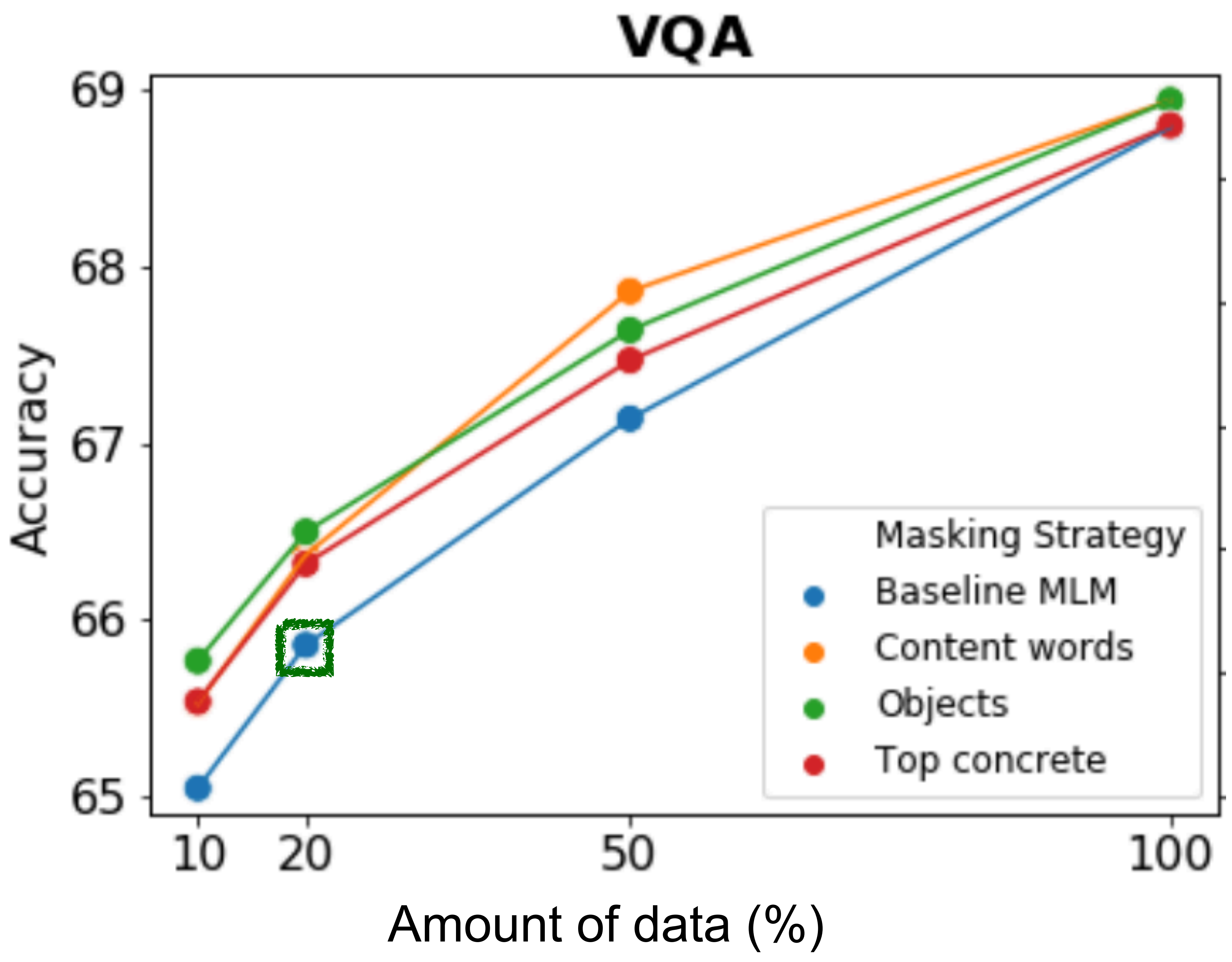


Data Efficient Training

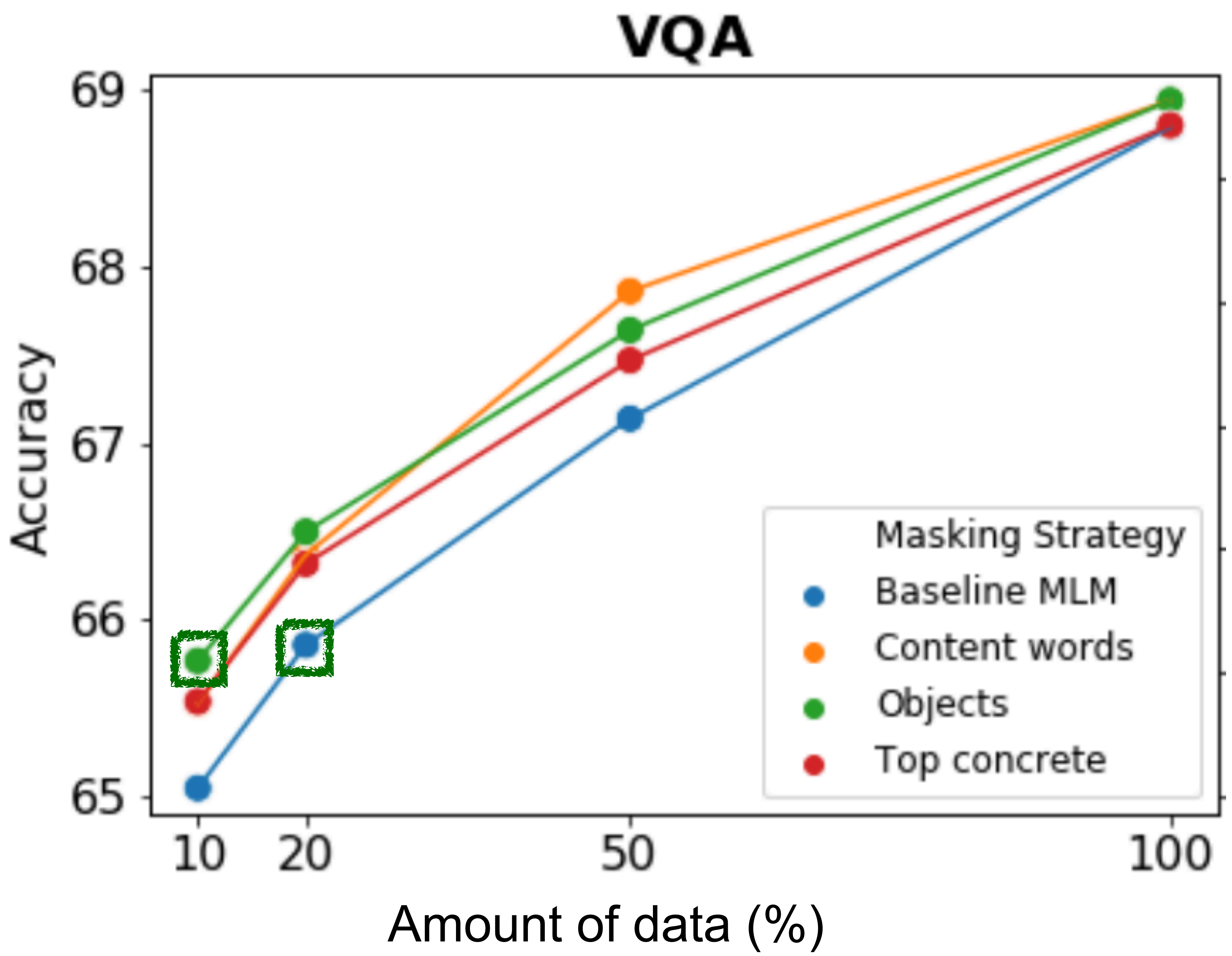


Amount of data (%)

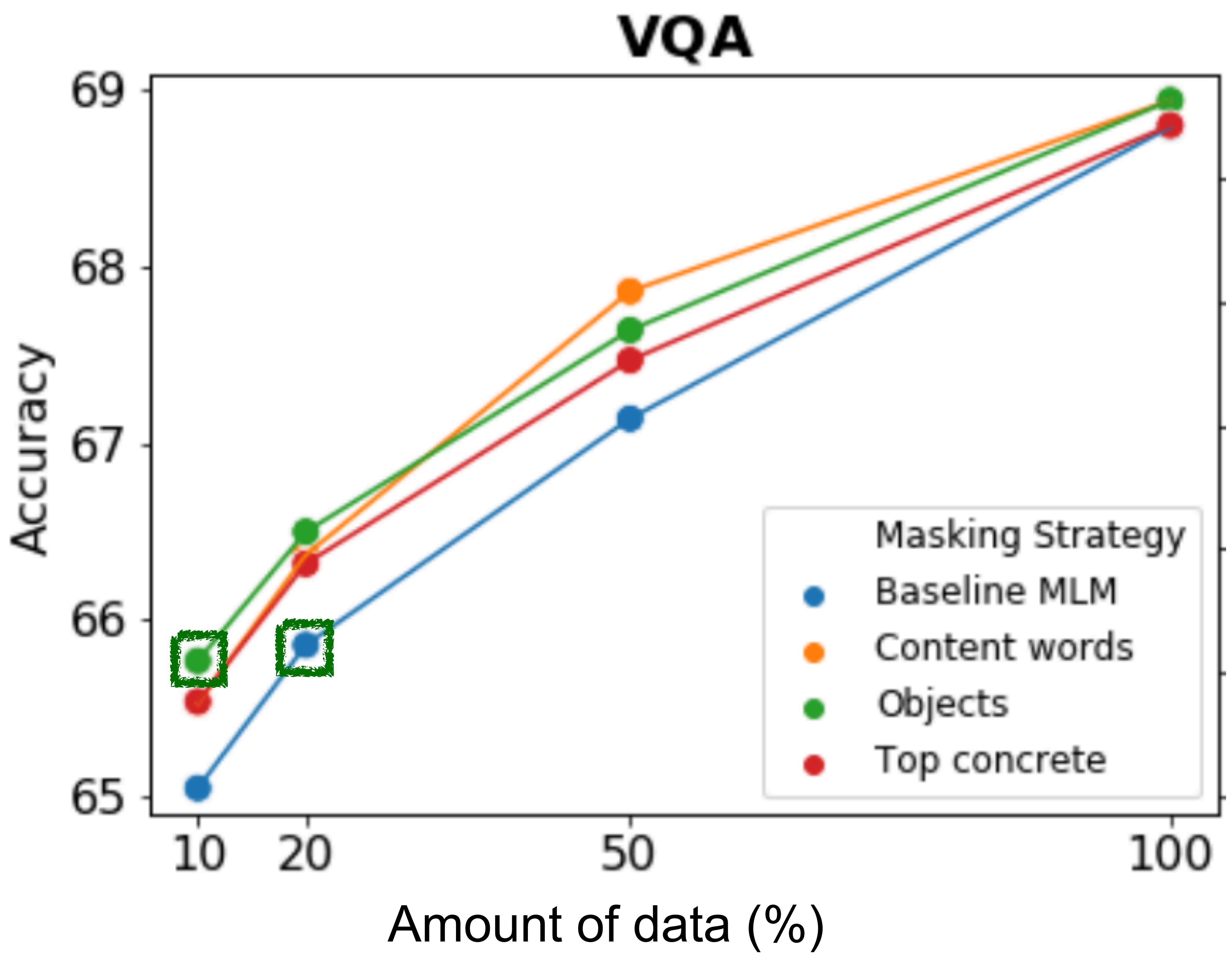
Data Efficient Training



Data Efficient Training



Data Efficient Training



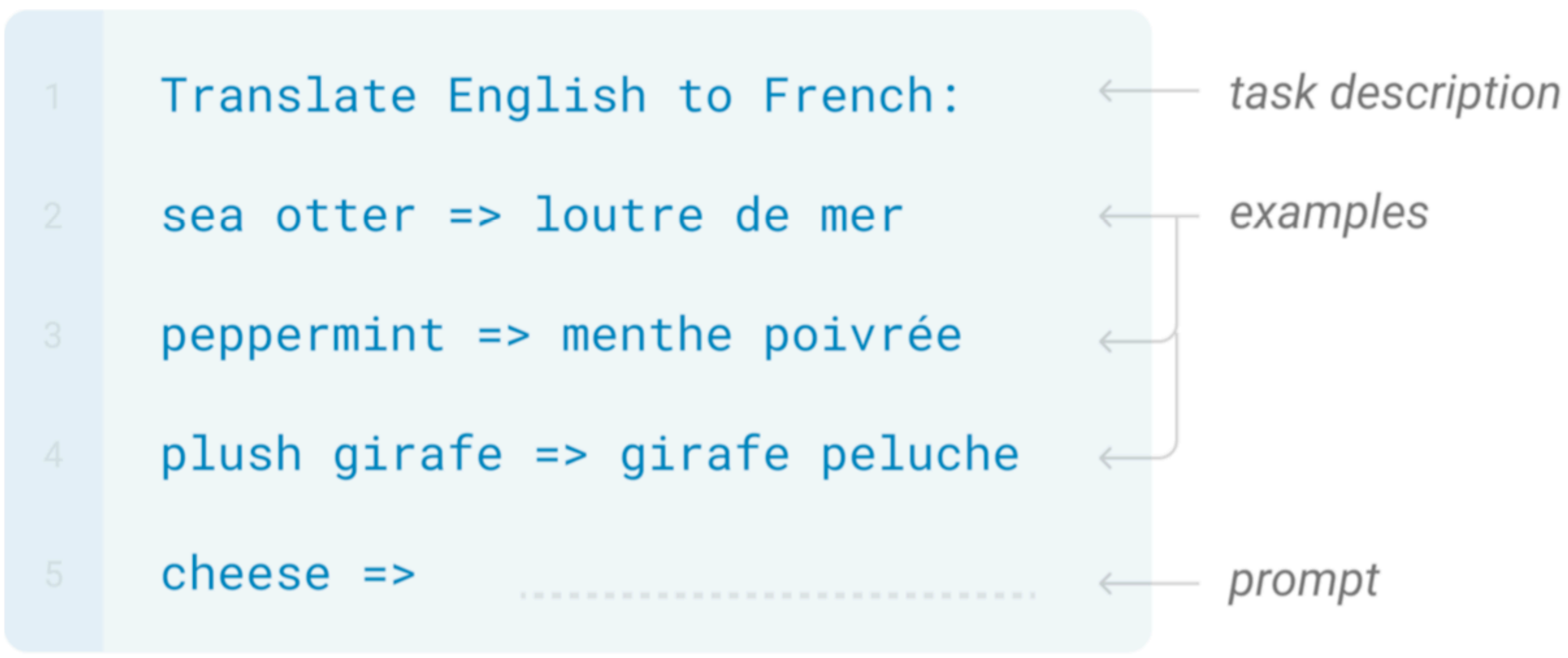
Similar accuracy, twice as fast

Few-shot Learning

- Only use a handful of examples to train a model

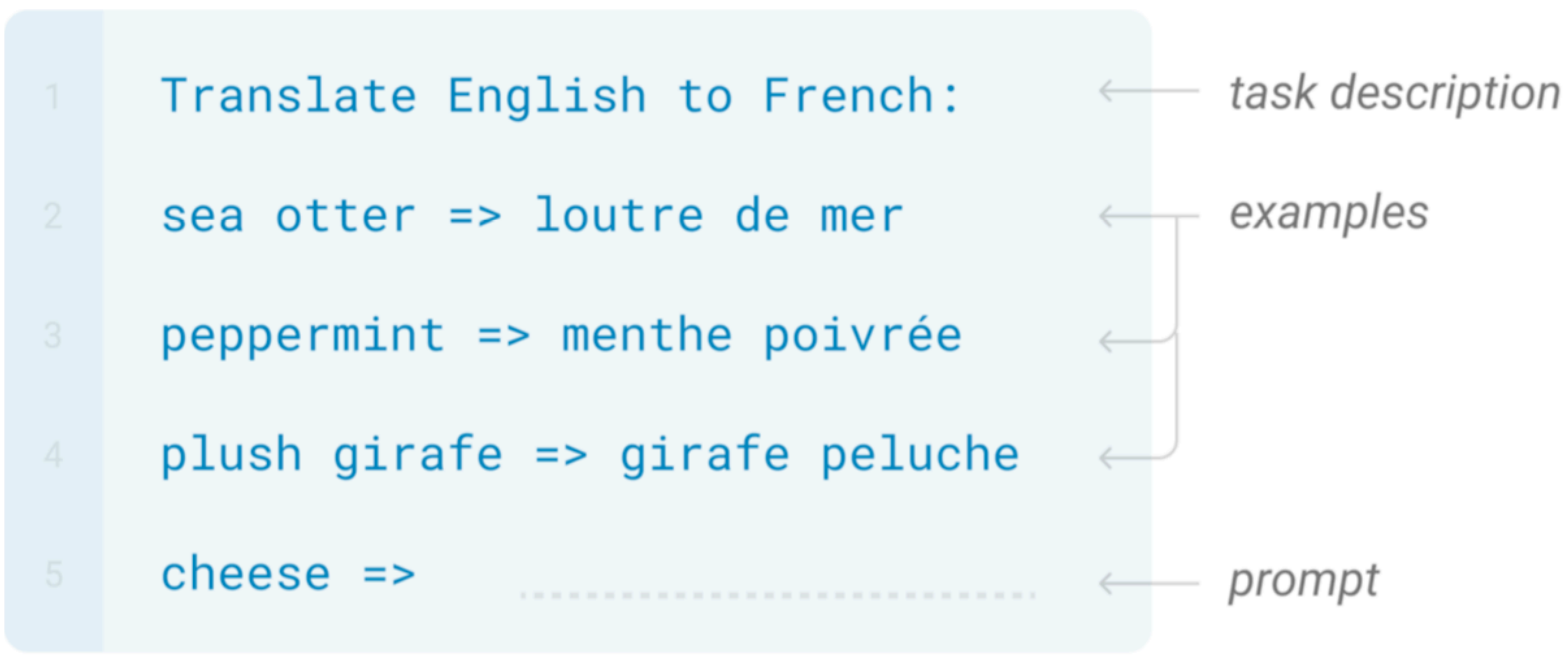
Few-shot Learning

- Only use a handful of examples to train a model
- Prompting
  - Brown et al. (2020), Schick & Sch¨utze, 2021)



Few-shot Learning

- Only use a handful of examples to train a model
- Prompting
  - Brown et al. (2020), Schick & Sch¨utze, 2021)
- Non-prompting methods
  - Mahabadi et al. (2022)

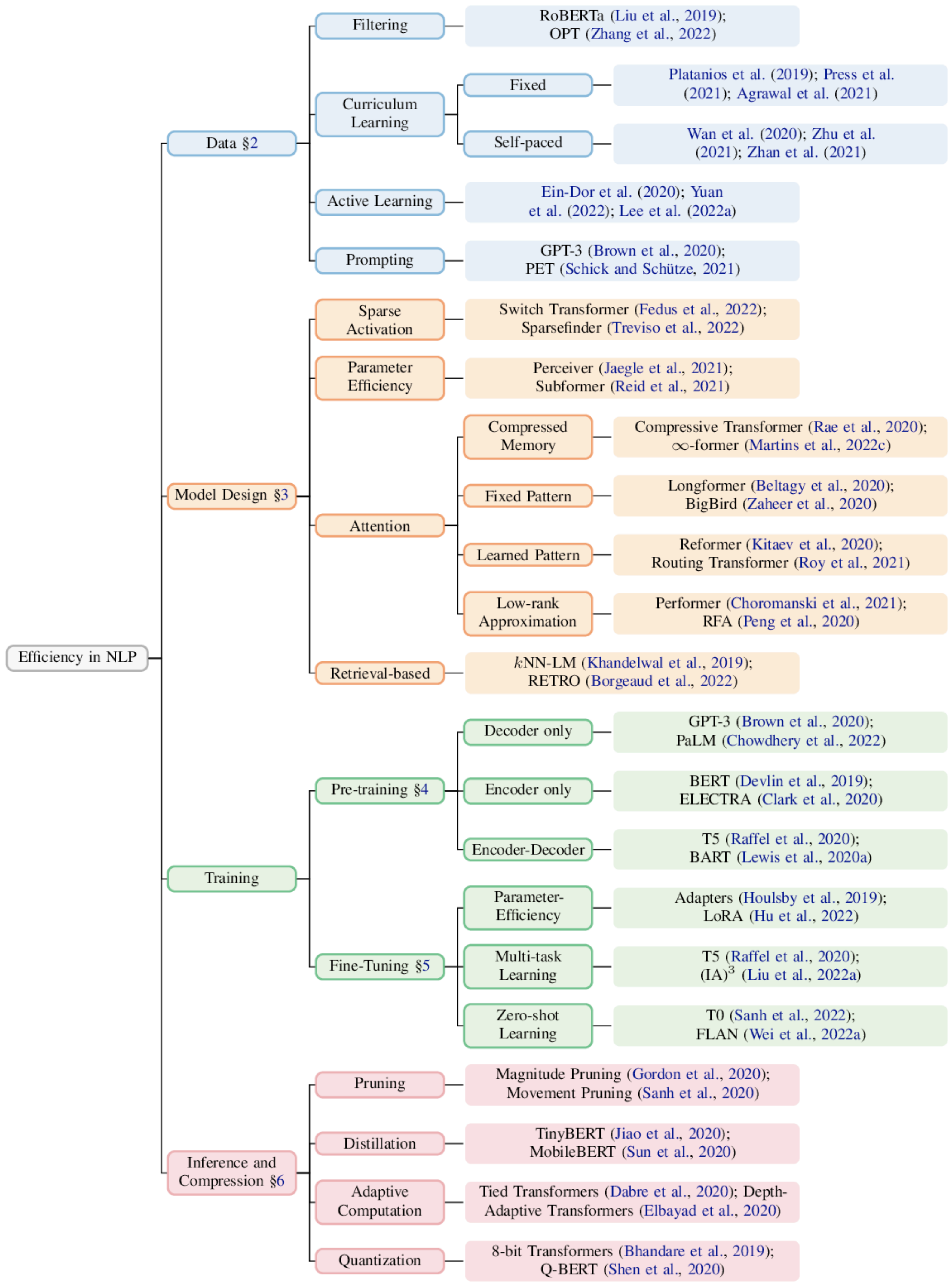


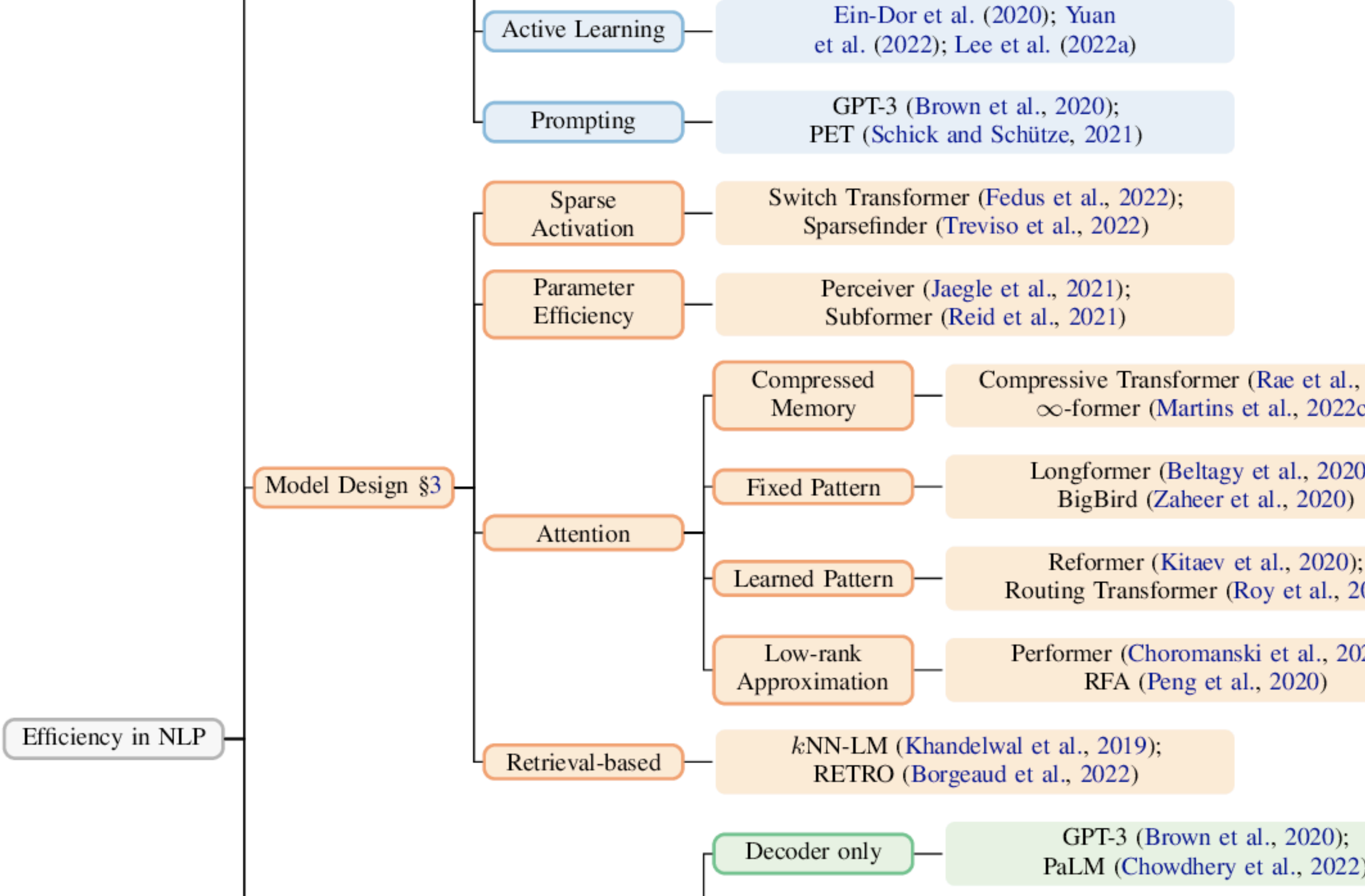
Data Efficiency

Open Questions

- Do we really need massive web-scale data to train our models?
    - Can we get along with less?
    - Sorscher et al. (2022)

Efficient Methods in NLP



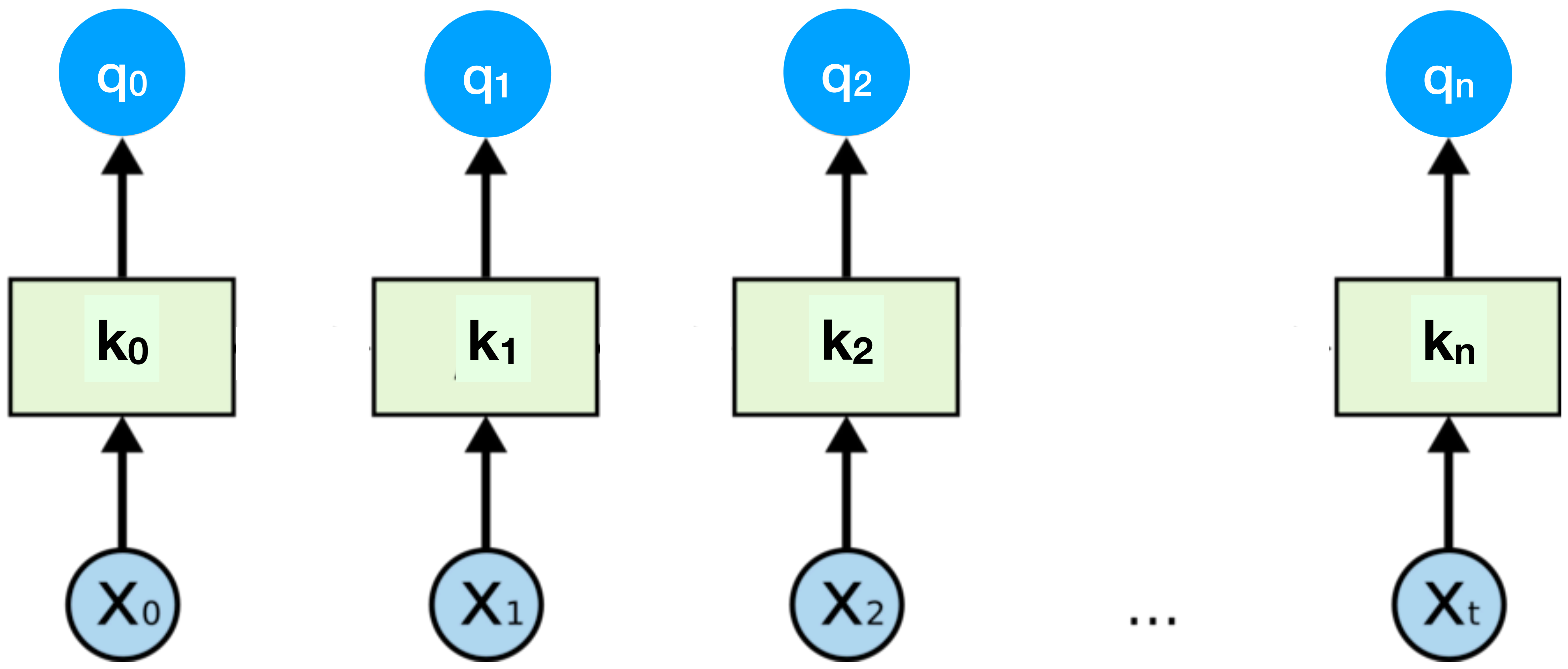


Active Learning
Ein-Dor et al. (2020); Yuan et al. (2022); Lee et al. (2022a)
Prompting
GPT-3 (Brown et al., 2020); PET (Schick and Schütze, 2021)
Sparse Activation
Switch Transformer (Fedus et al., 2022); Sparsefinder (Treviso et al., 2022)
Parameter Efficiency
Perceiver (Jaegle et al., 2021); Subformer (Reid et al., 2021)
Compressed Memory
Compressive Transformer (Rae et al.,
∞-former (Martins et al.,
Fixed Pattern
Longformer (Beltagy et al.,
BigBird (Zaheer et al., 2020)
Learned Pattern
Reformer (Kitaev et al., 2020);
Routing Transformer (Roy et al.,
Low-rank Approximation
Performer (Choromanski et al.,
RFA (Peng et al., 2020)
Model Design §3
Attention
Efficiency in NLP
Retrieval-based
kNN-LM (Khandelwal et al., 2019); RETRO (Borgeaud et al., 2022)
Decoder only
GPT-3 (Brown et al., 2020);
PaLM (Chowdhery et al., 2022

Transformers

Vaswani et al., 2017

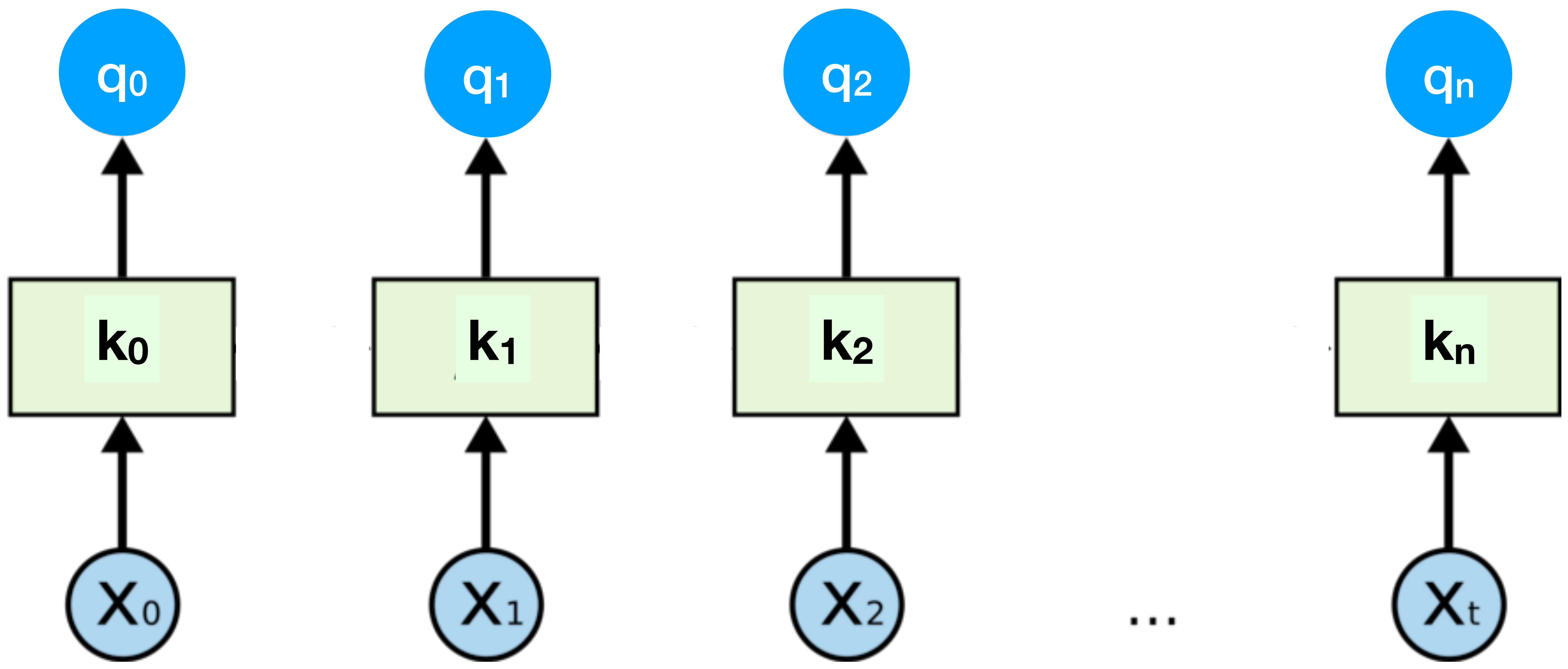
- **The** method for text representation
  - Also for vision, speech, combio, …



Transformers

Vaswani et al., 2017

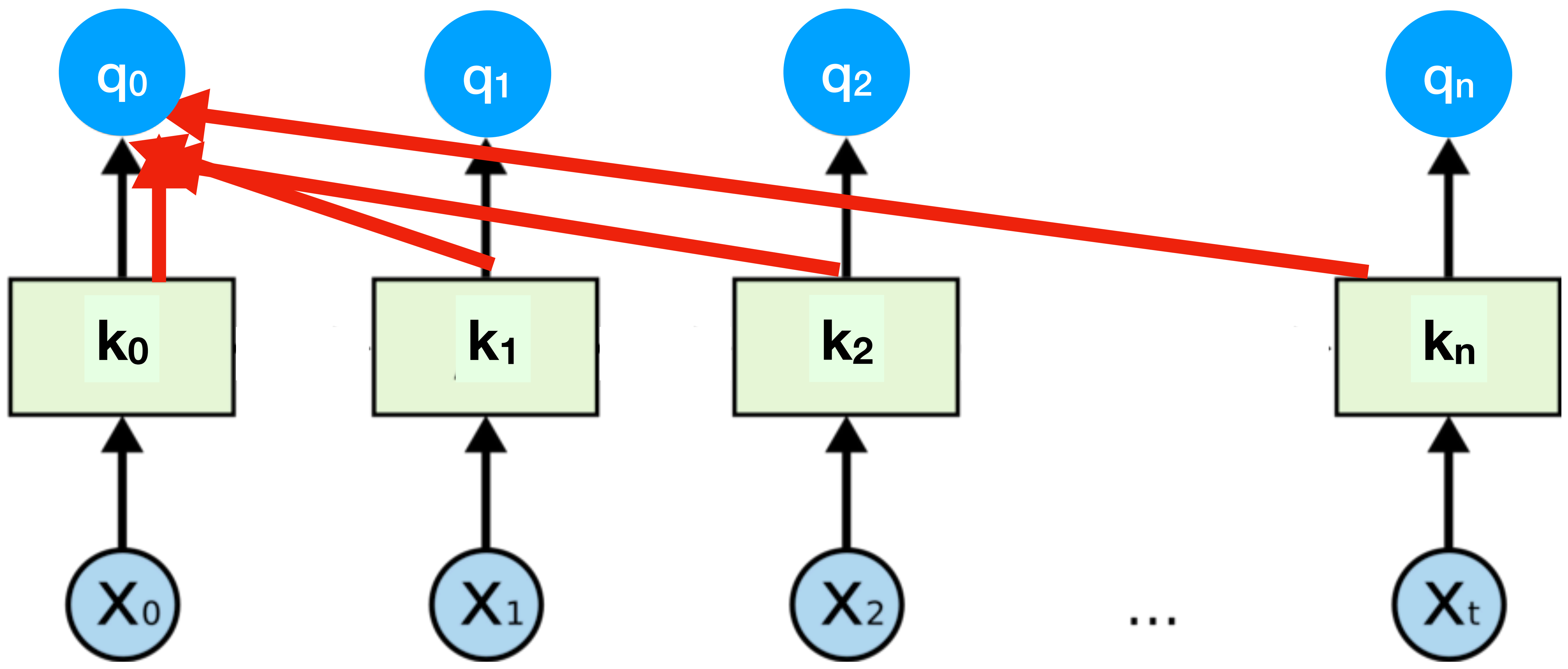
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

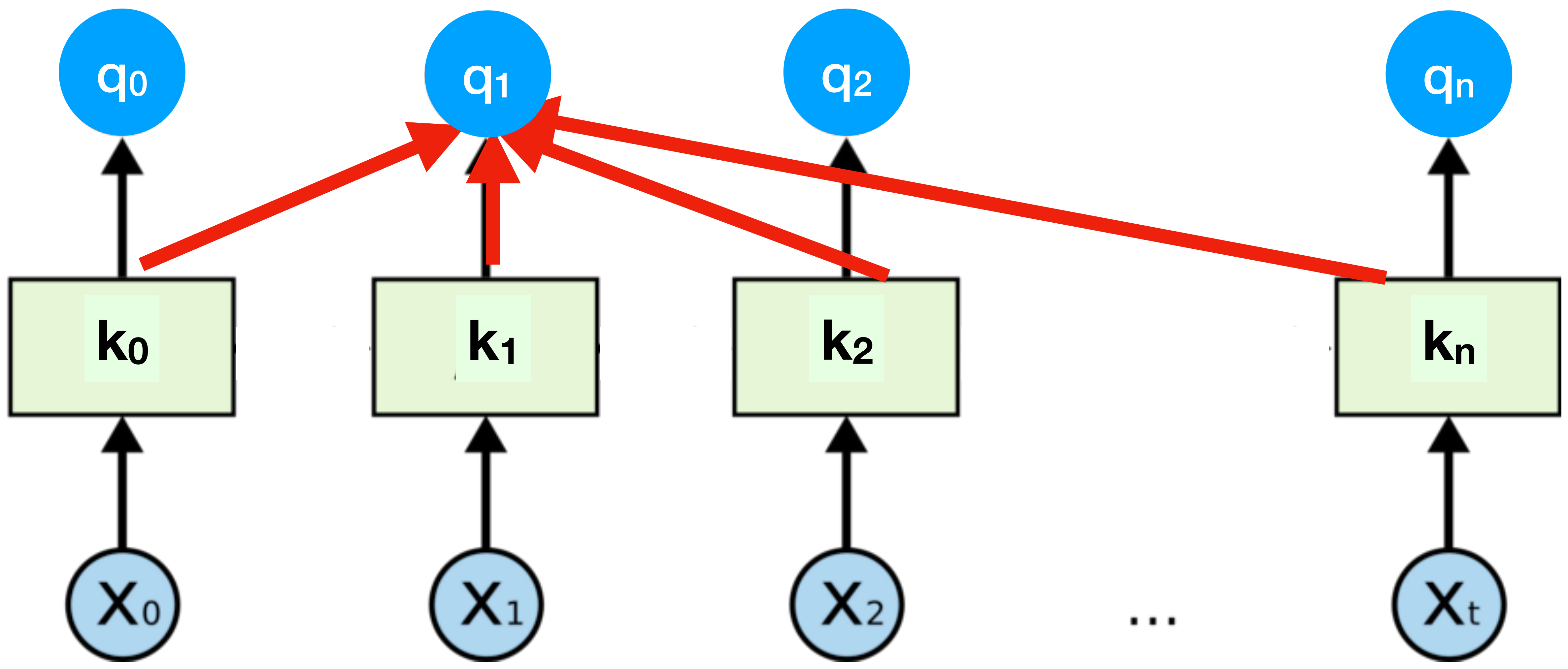
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

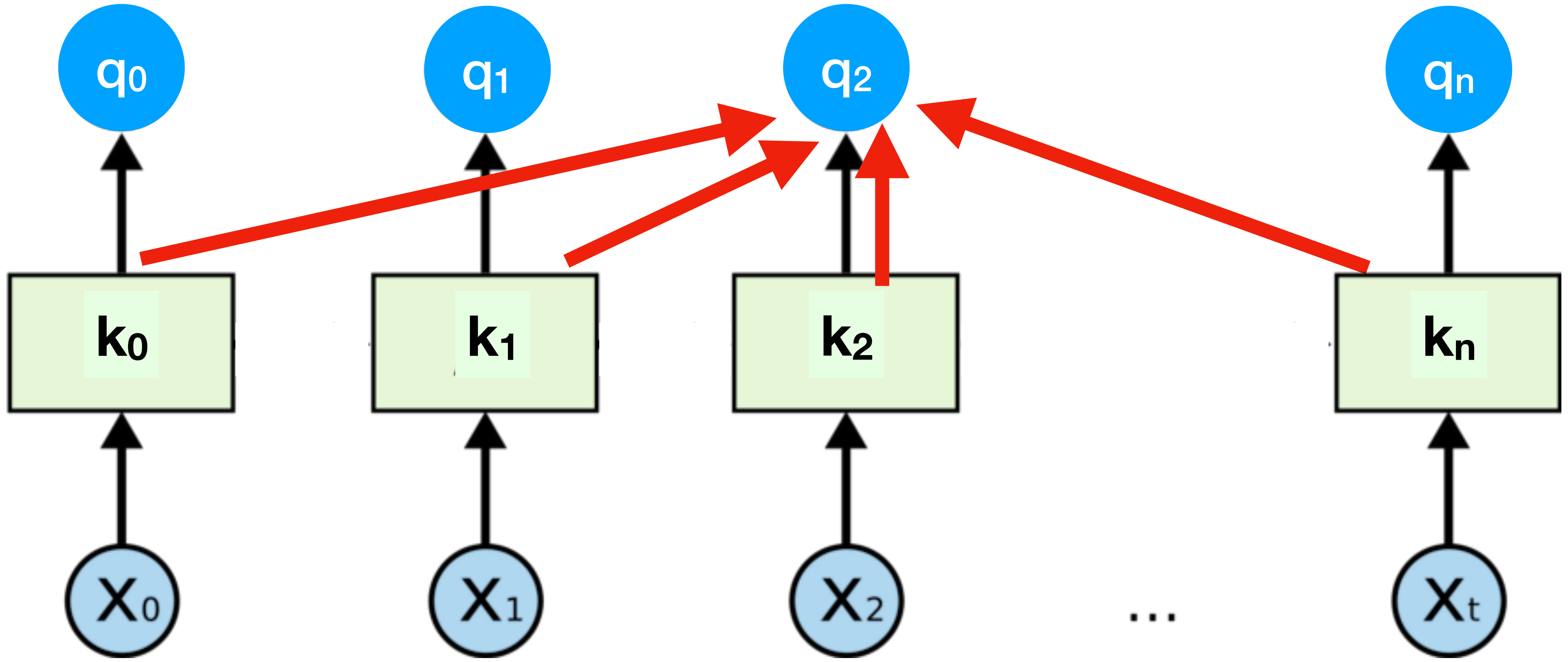
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

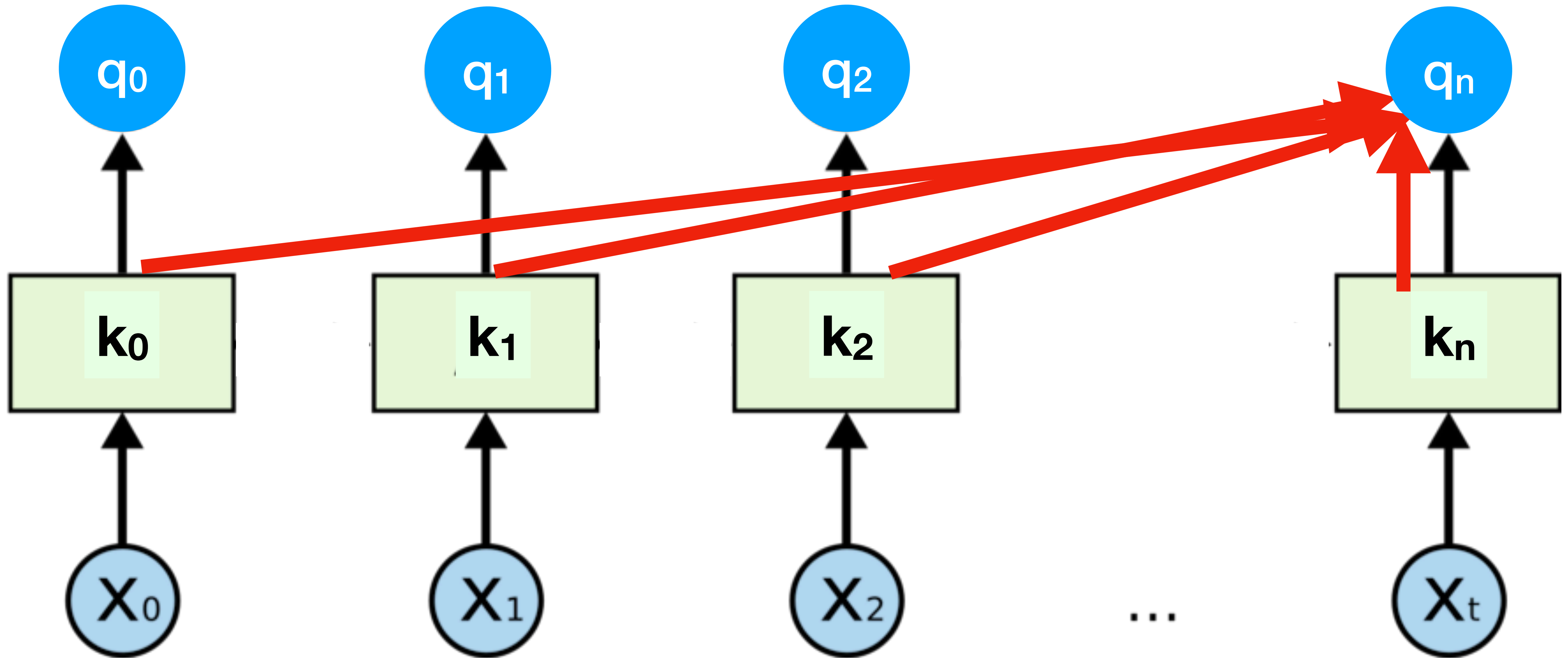
- **The** method for text representation
    - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

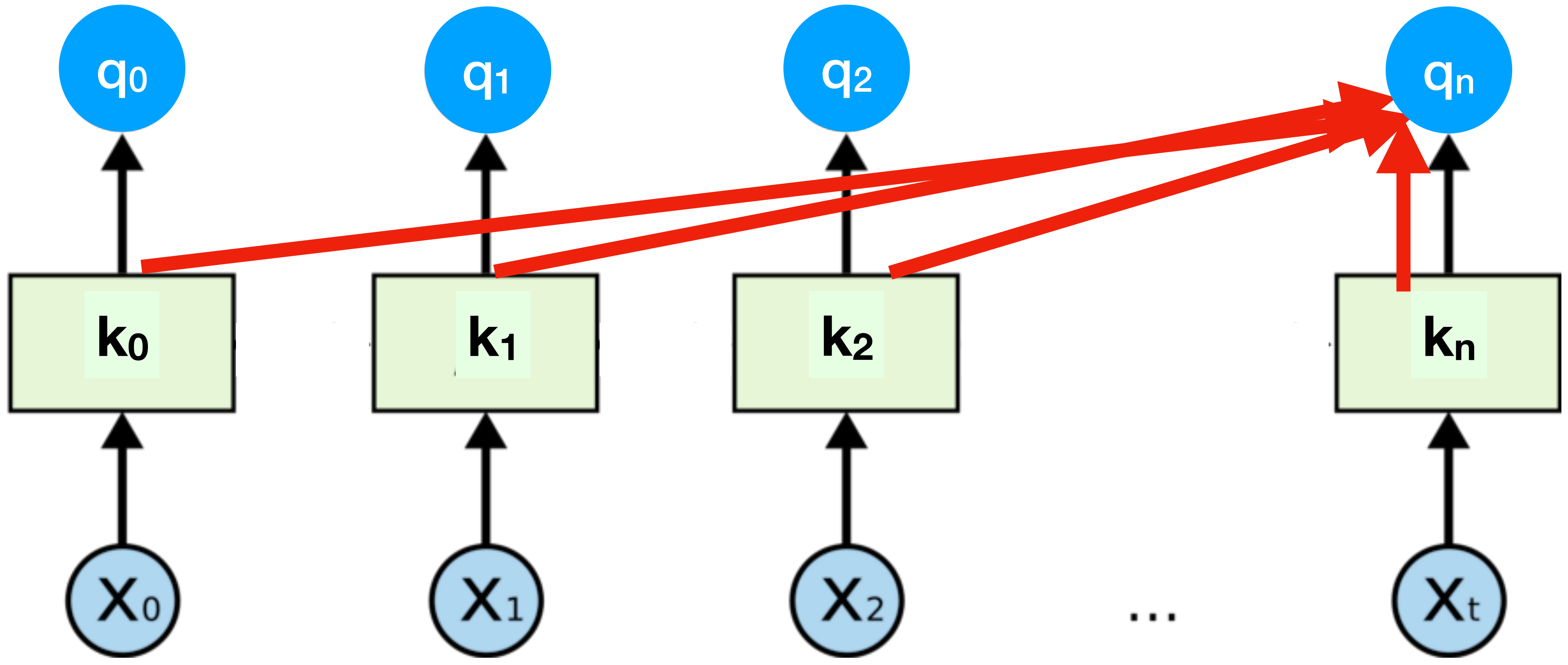
- **The** method for text representation
    - Also for vision, speech, combio, …
- Each word attends to all other words



Transformers

Vaswani et al., 2017

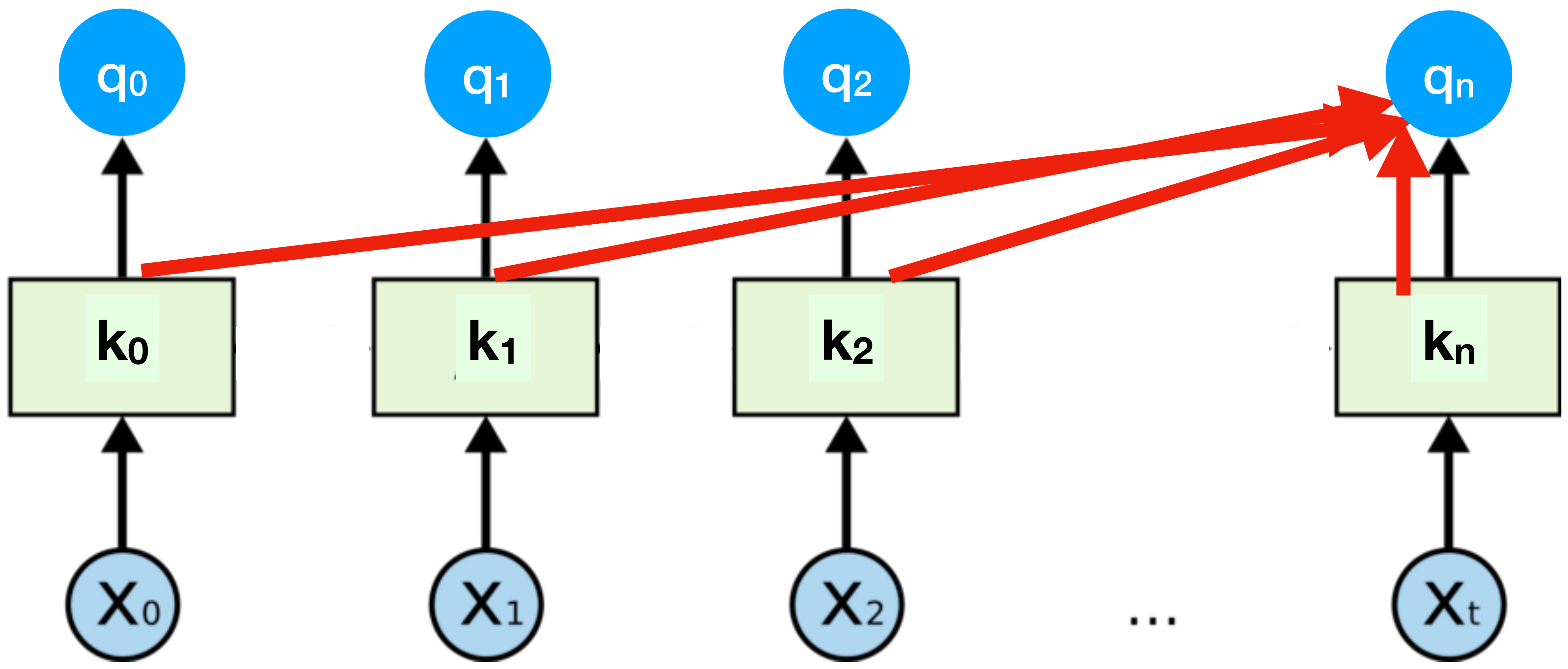
- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words
- $O(n^2)$ complexity in the sentence length n



Transformers

Vaswani et al., 2017

- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word attends to all other words
- O(n^2) complexity in the sentence length n
- Fatal for long sequences
  - Books, articles, etc.

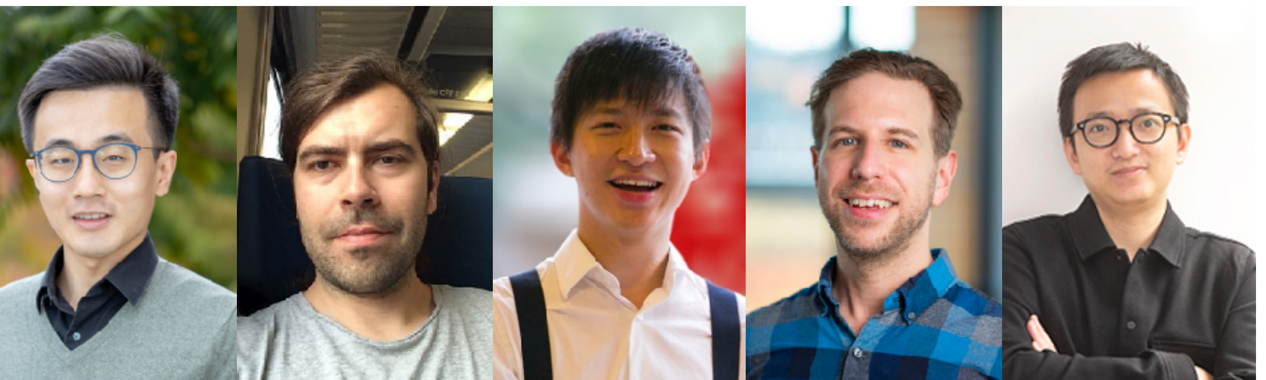


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea:** approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)

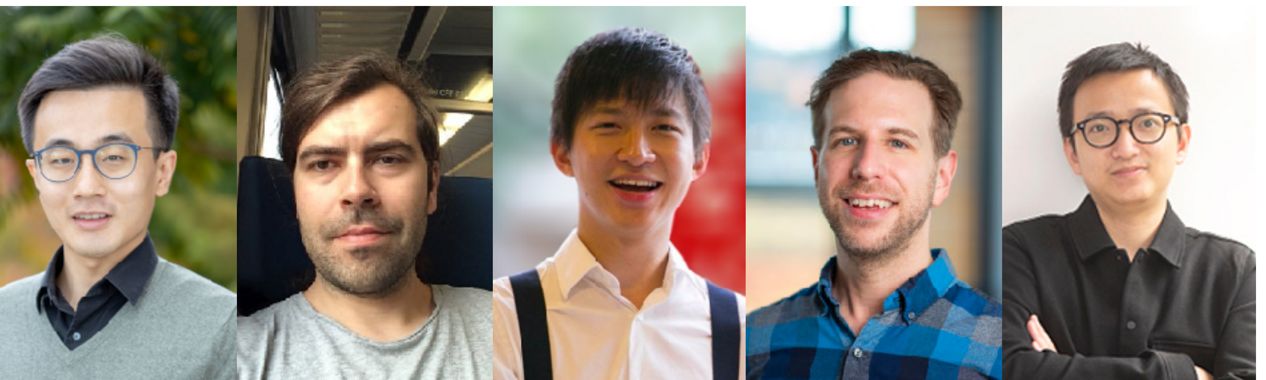


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea:** approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math

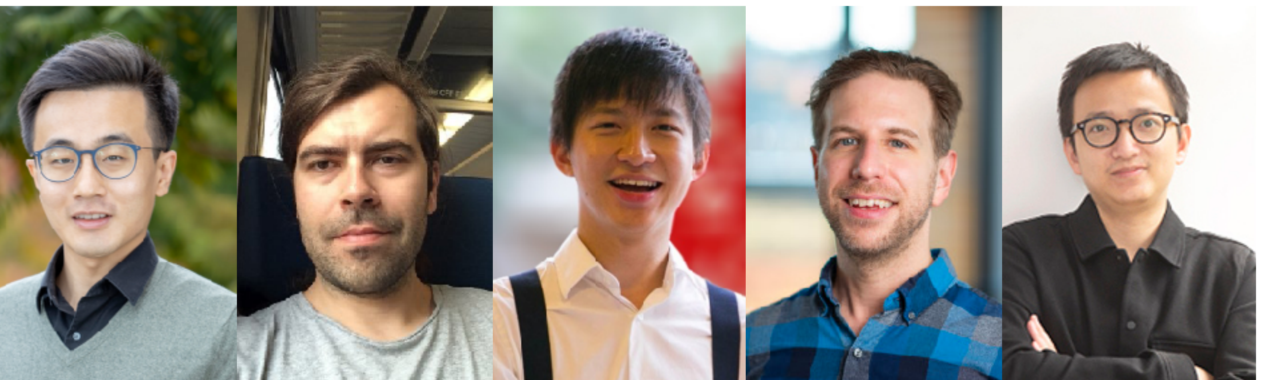


Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

- **Key idea:** approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math

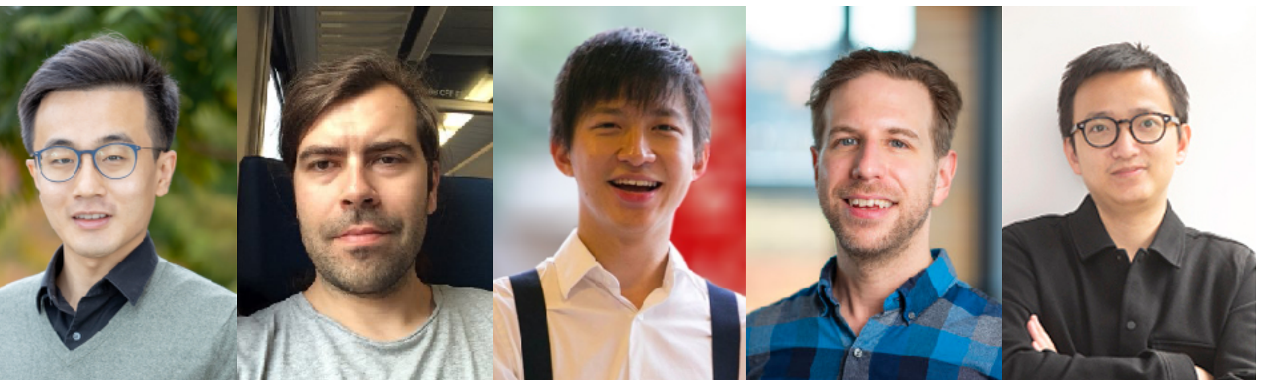


Random Feature Attention

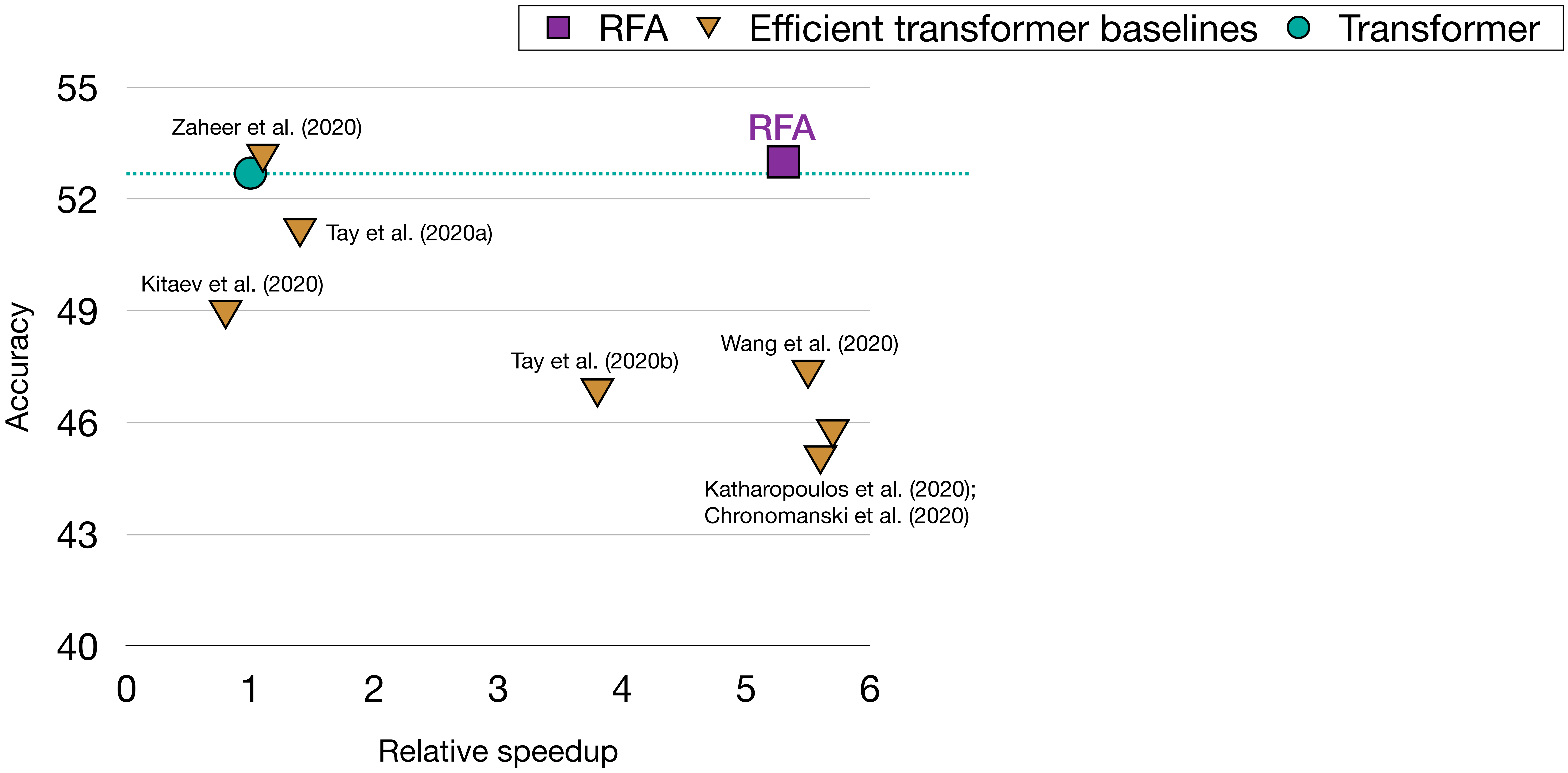
Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

spotlight presentation

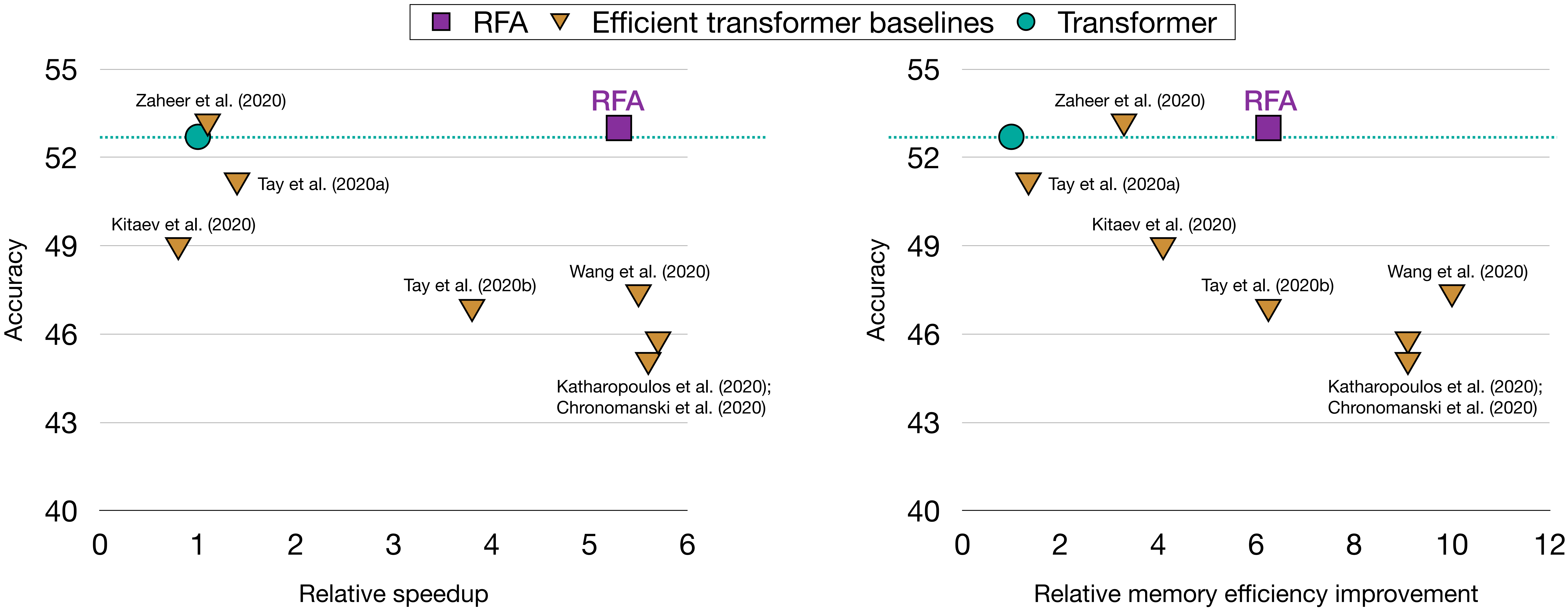
- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math
- Linear runtime and memory requirements



Better Efficiency-Accuracy Tradeoff



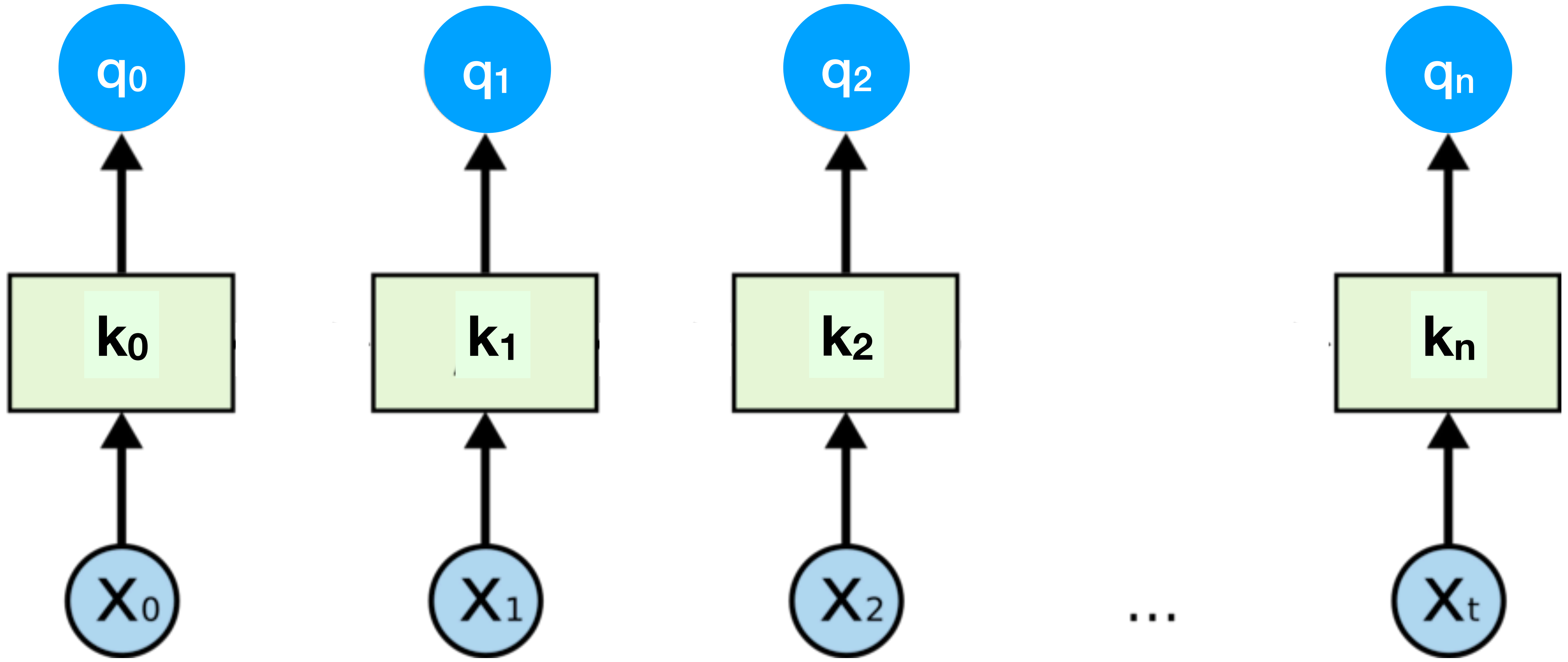
Better Efficiency-Accuracy Tradeoff

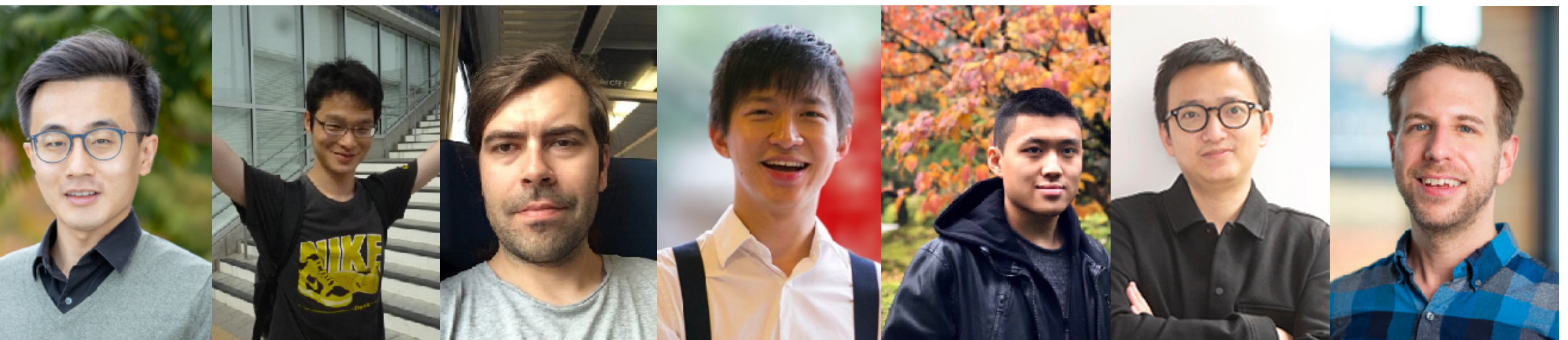


ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

- **Key intuition**: treat the sentence as **memory of size *n***

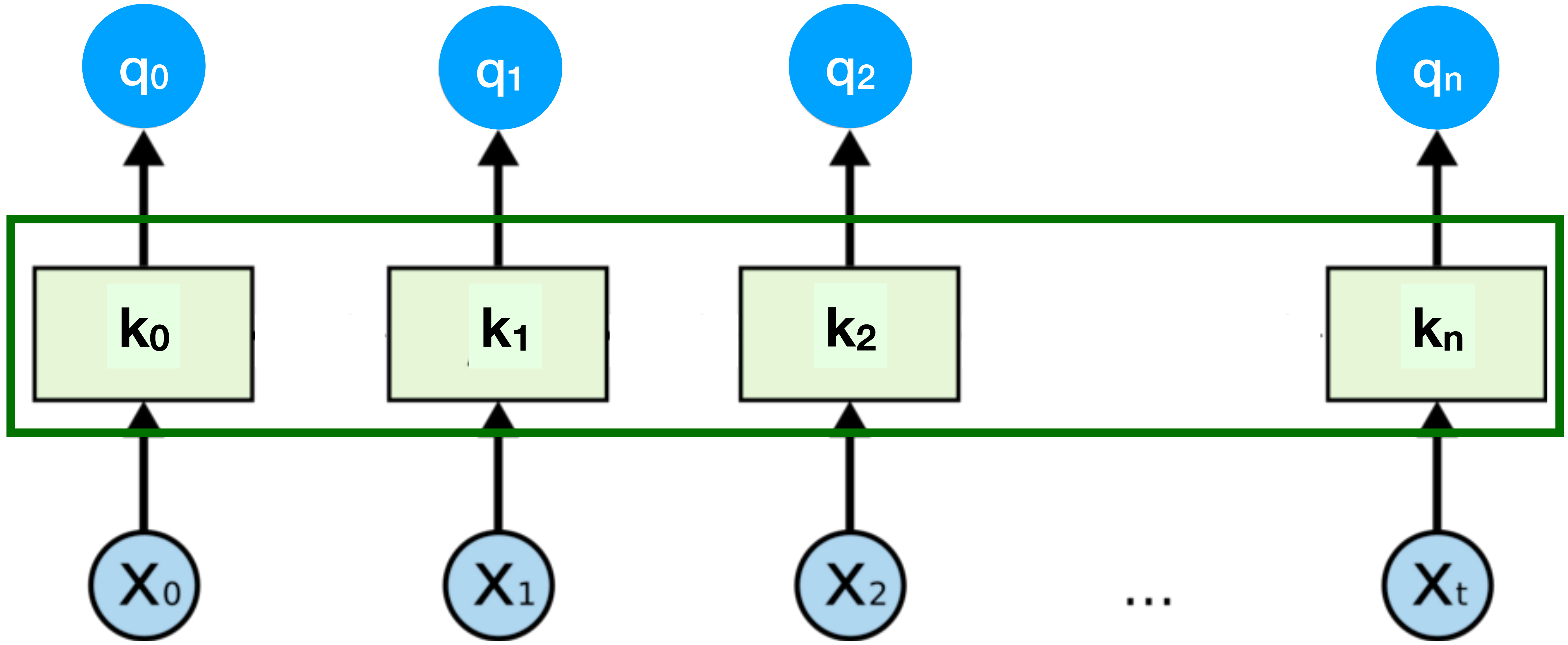


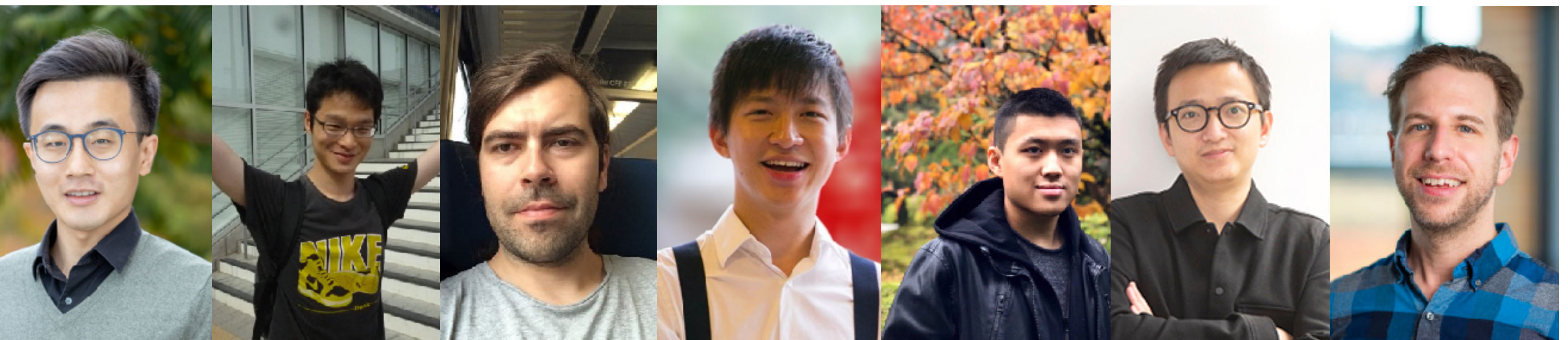


ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

- **Key intuition**: treat the sentence as **memory of size *n***

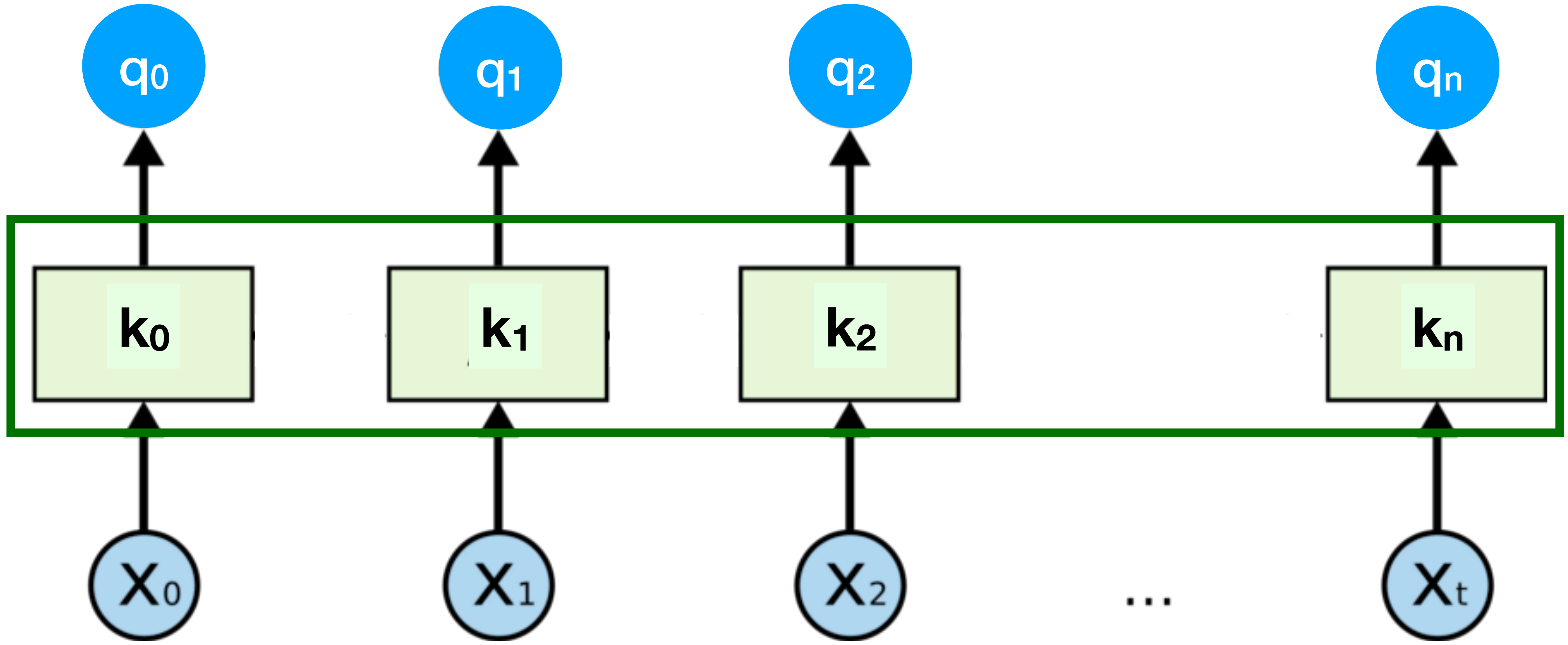


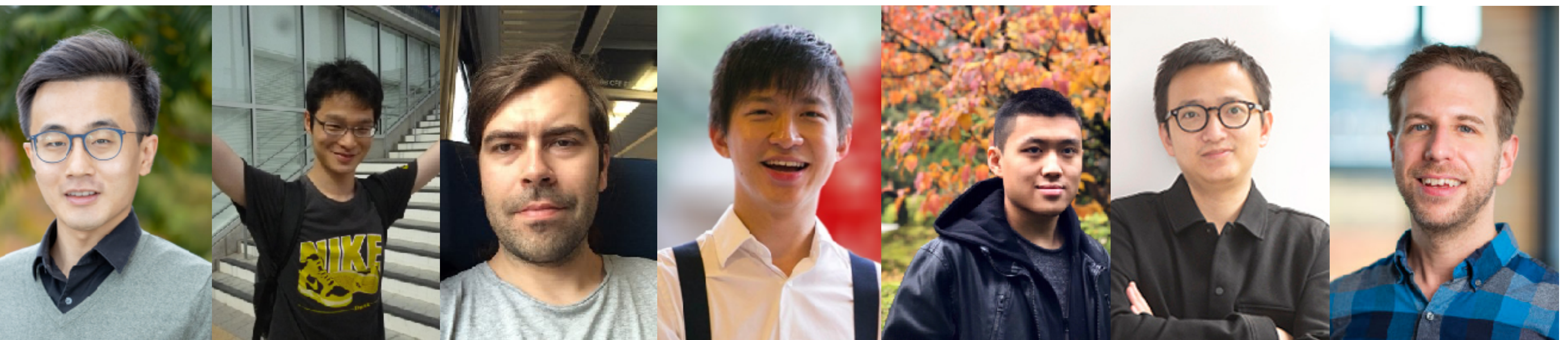


ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

- **Key intuition**: treat the sentence as **memory of size *n***
- **Key idea**: replace this memory with a fixed size memory of (fixed) size $k << n$
  - Instead of attending n tokens, each word attends to k tokens

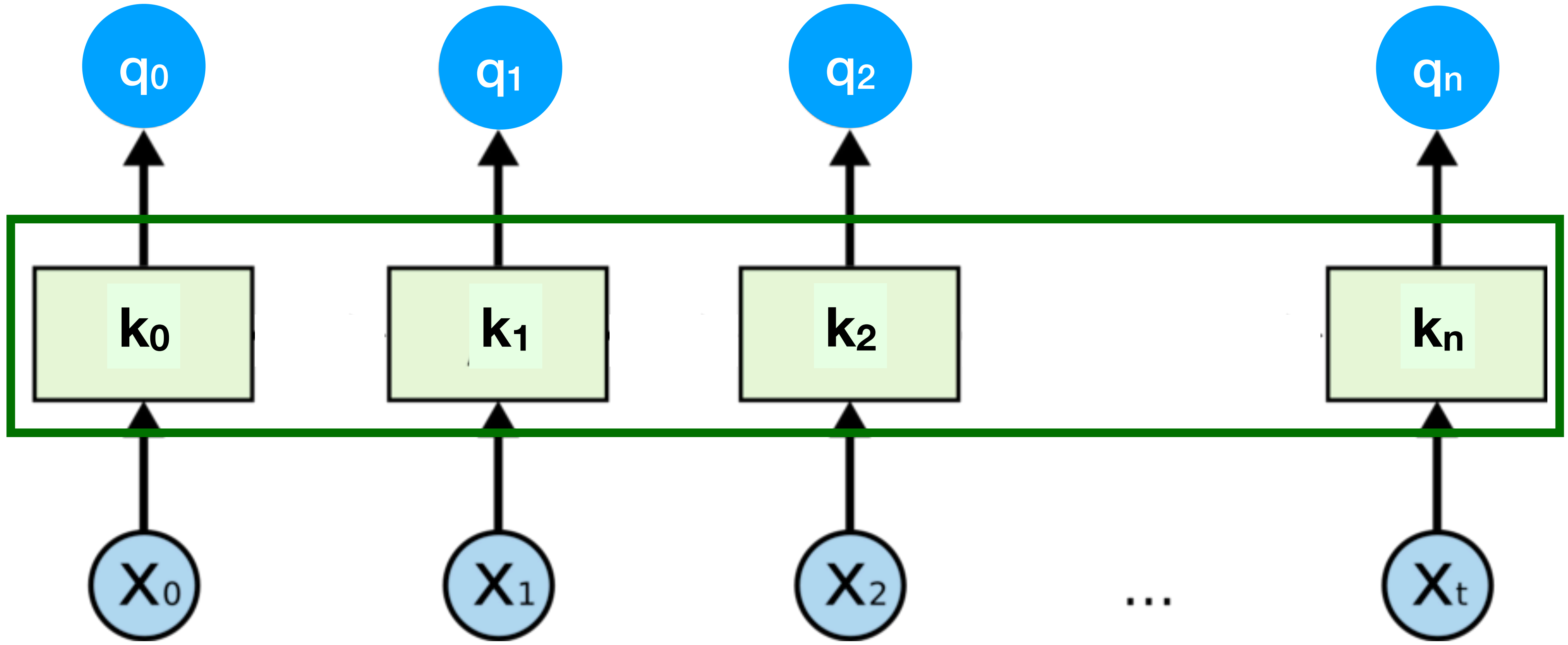


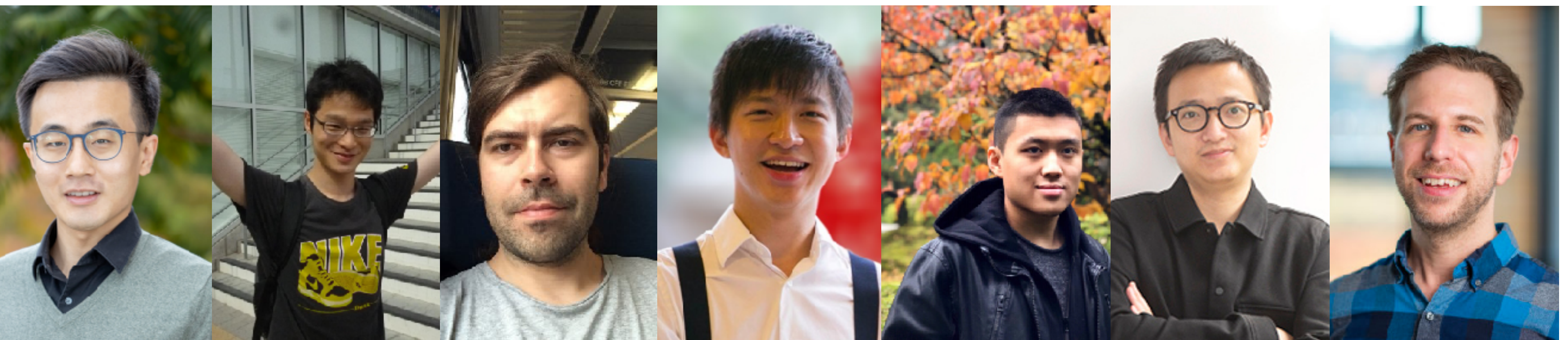


ABC: Attention with Bounded-memory Control

Peng, Kasai, Pappas, Yogatama, Wu, Kong, S. & Smith, ACL 2022

- **Key intuition**: treat the sentence as **memory of size *n***
- **Key idea**: replace this memory with a fixed size memory of (fixed) size $k << n$
  - Instead of attending n tokens, each word attends to k tokens
- Overall complexity linear in n
  - With constant k





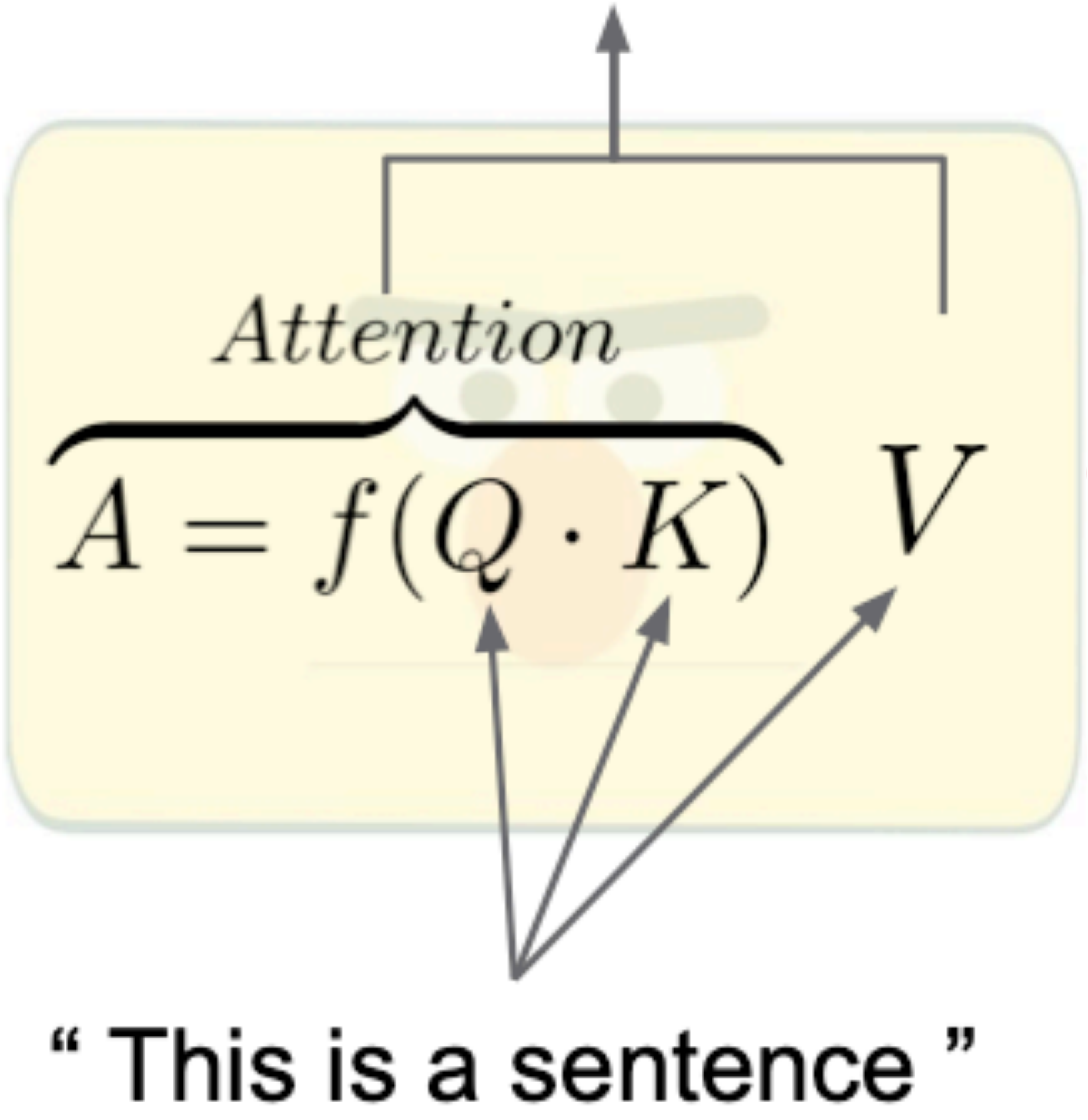
ABC Results

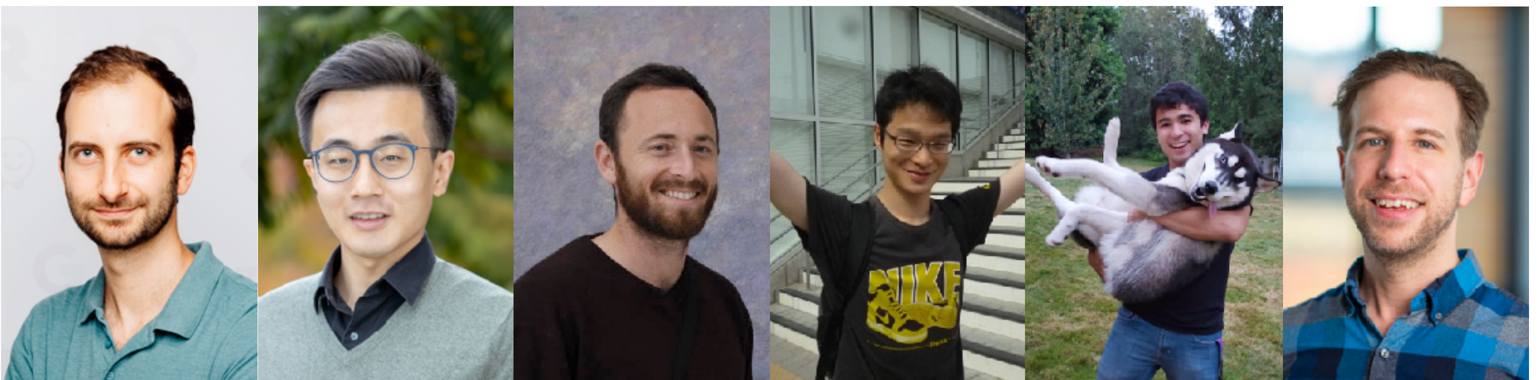
Speed

Memory

How Much Does Attention Actually Attend?

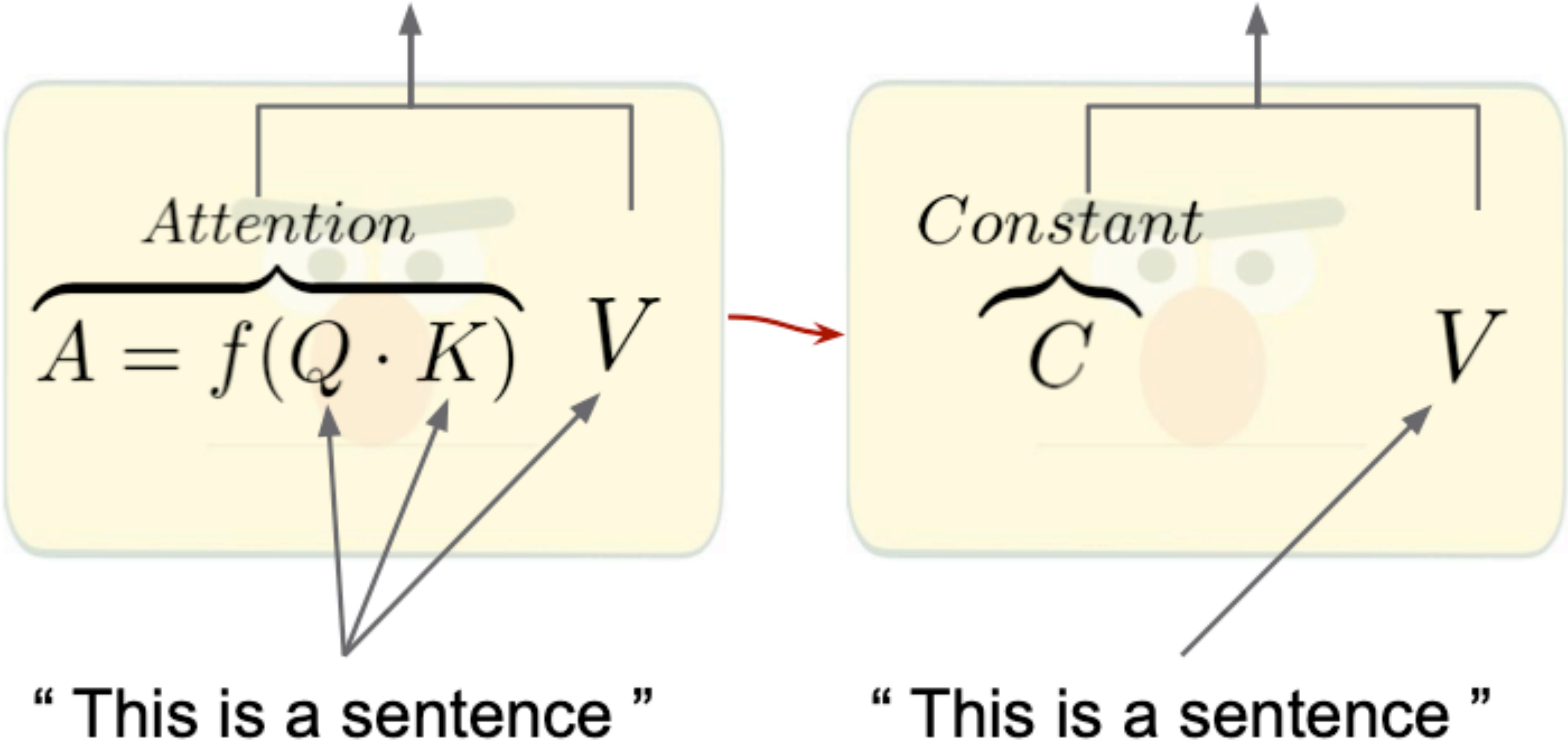
Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

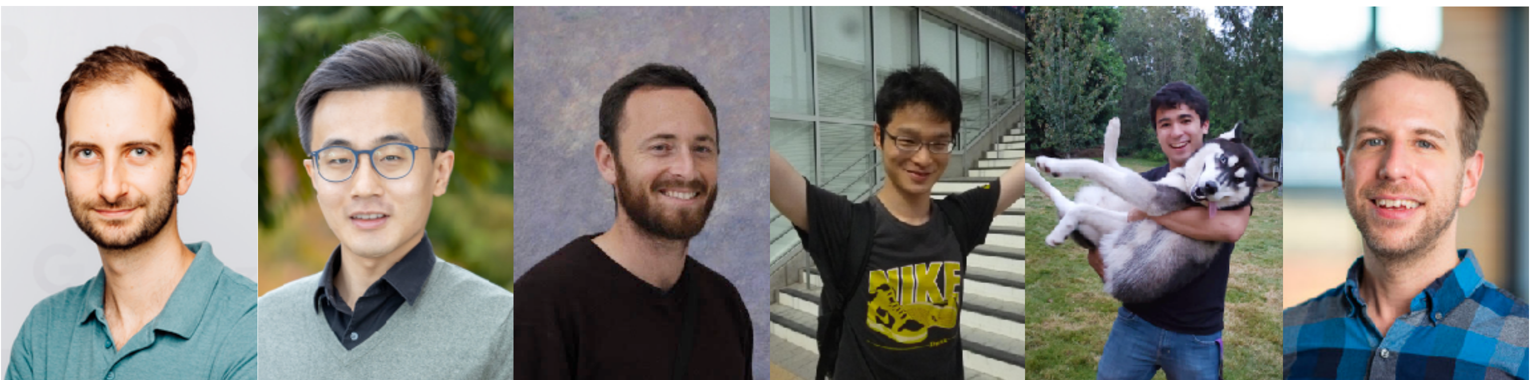




How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

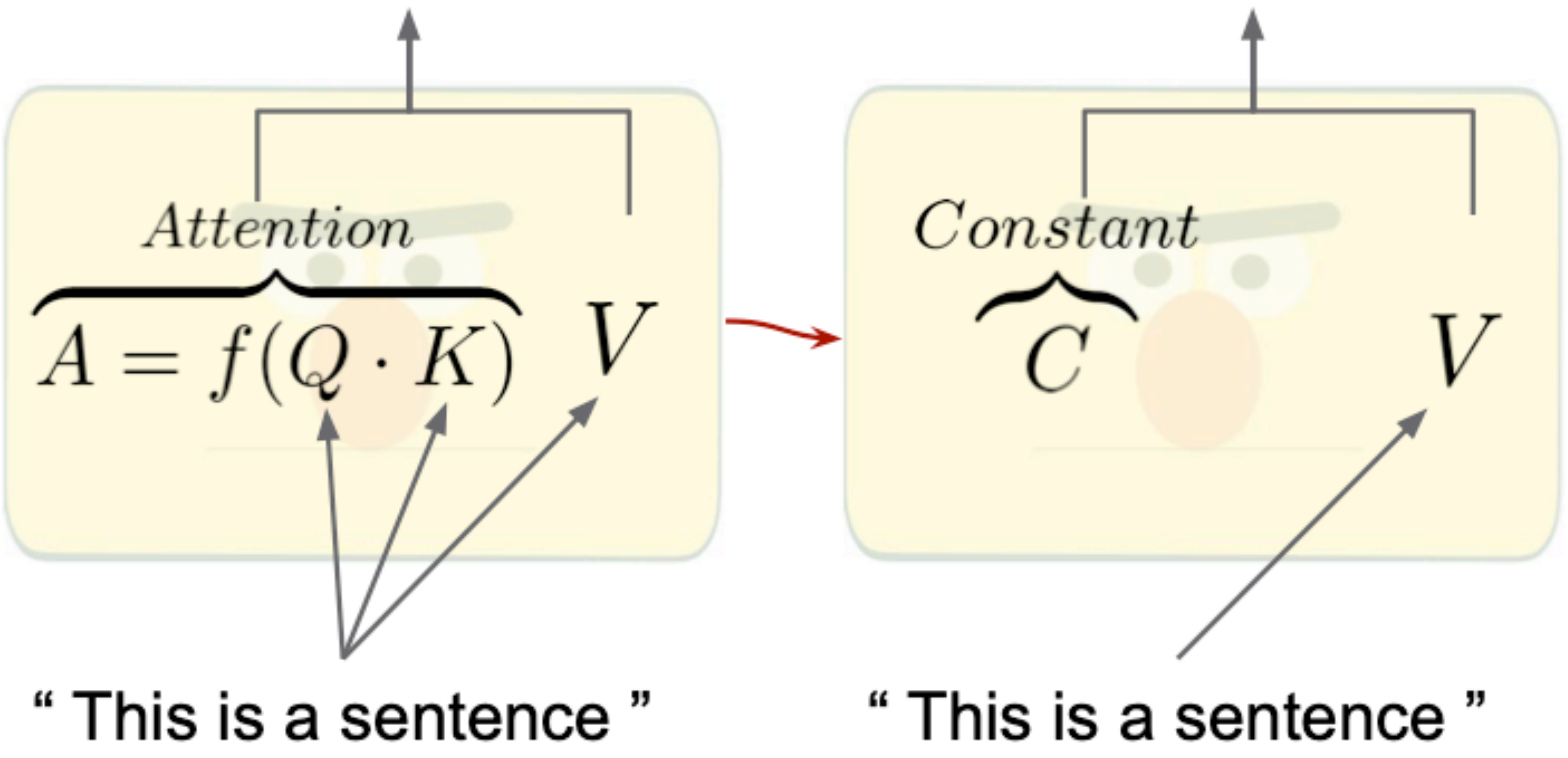


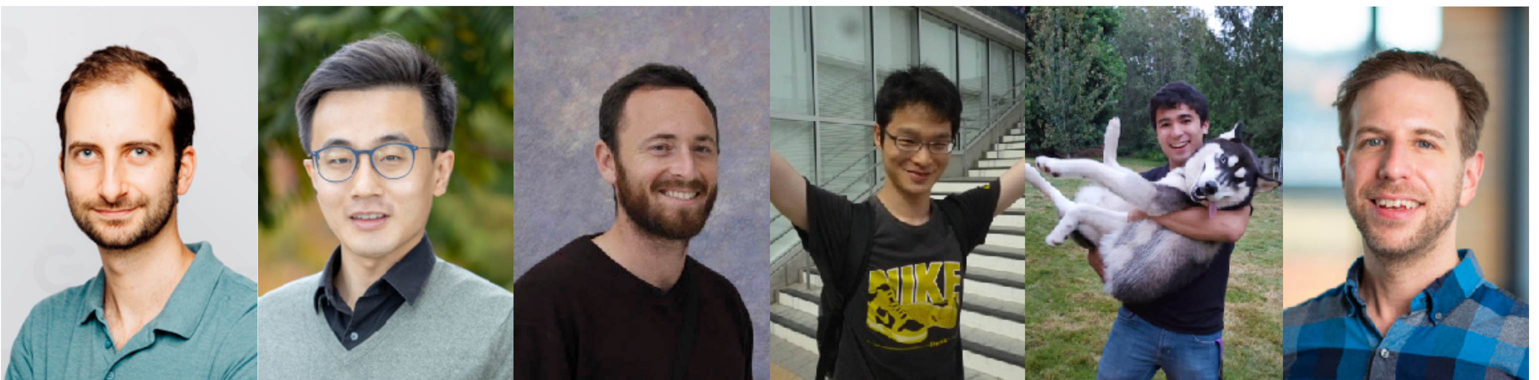


How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

- Model doesn't collapse
  - Average accuracy loss of 8% only

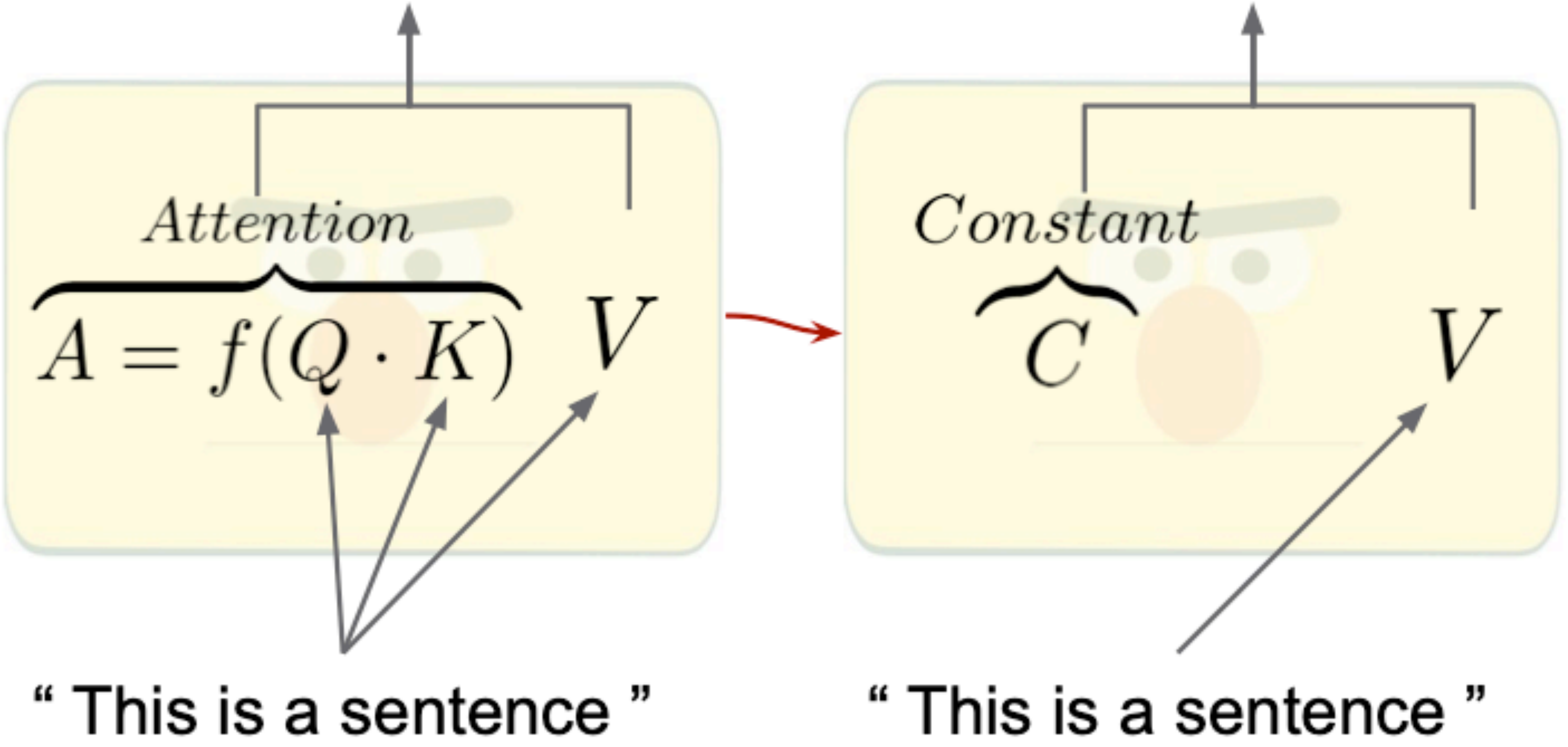


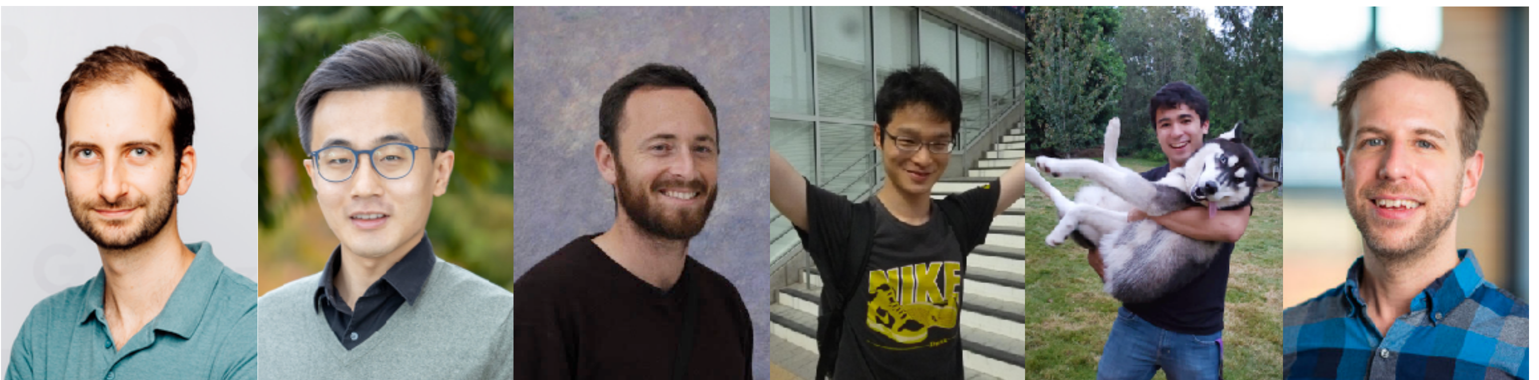


How Much Does Attention Actually Attend?

Hassid, Peng, Rotem, Kasai, Montero, Smith & S., Findings of EMNLP 2022

- Model doesn't collapse
  - Average accuracy loss of 8% only
- Potential for huge savings



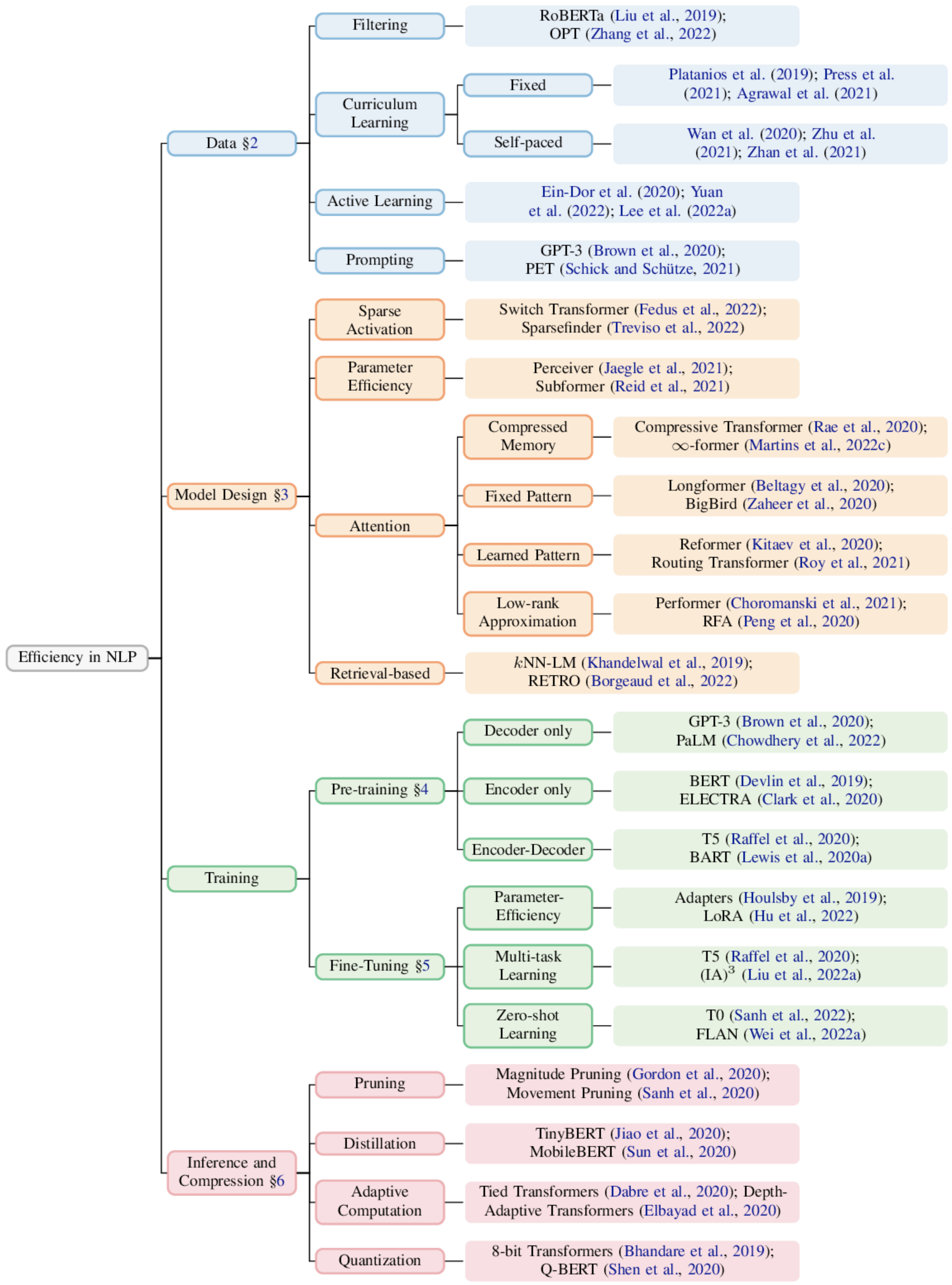


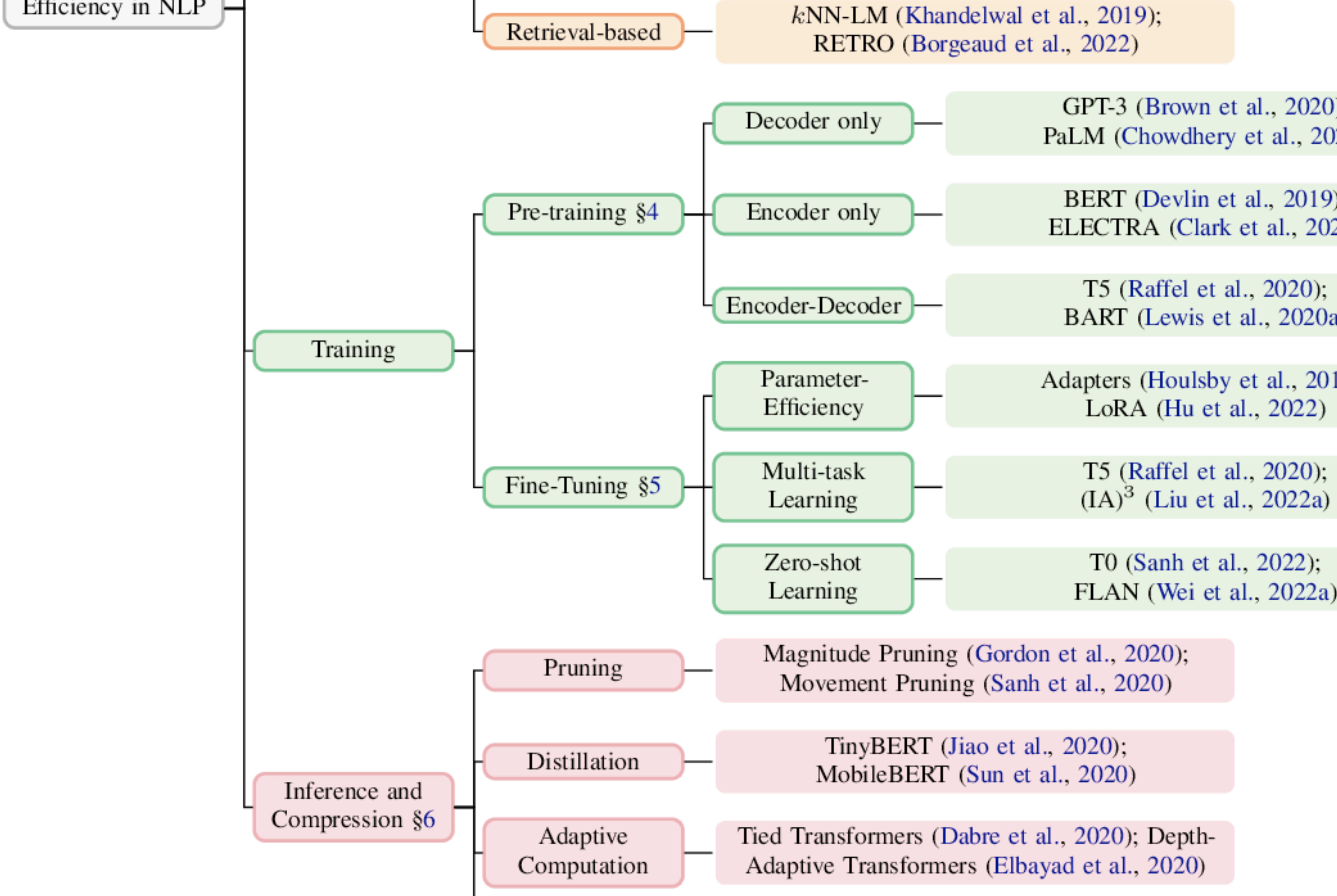
Efficient Modeling

Open Questions

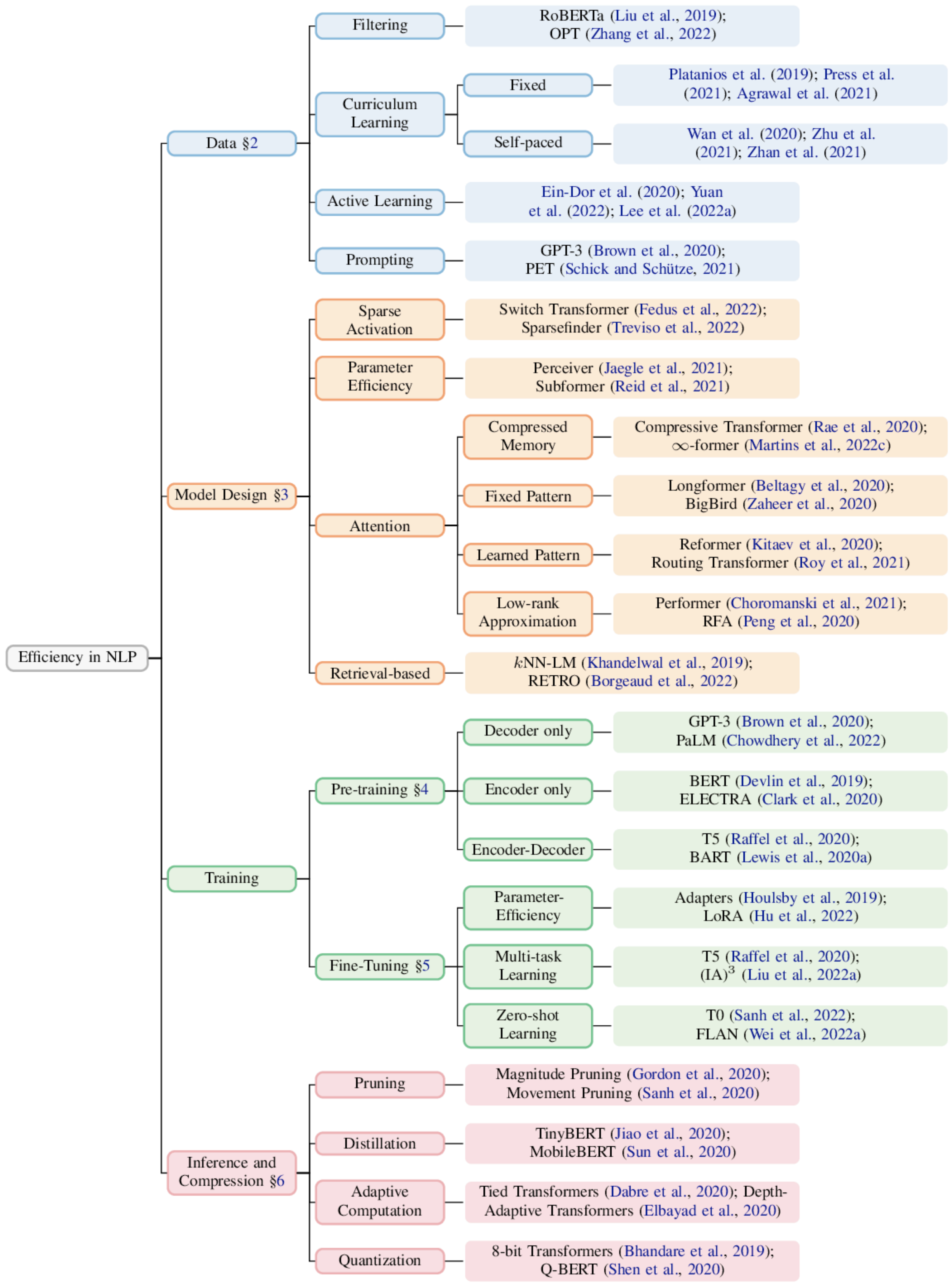
- Can we find the next generation of Transformers?
  - S4 (Gu et al., 2021)
- Should we store knowledge in the model parameters?
  - Retrieval-based models
  - Gu et al (2018); Lewis et al. (2020); Li et al. (2022); Borgeaud et al. (2022)

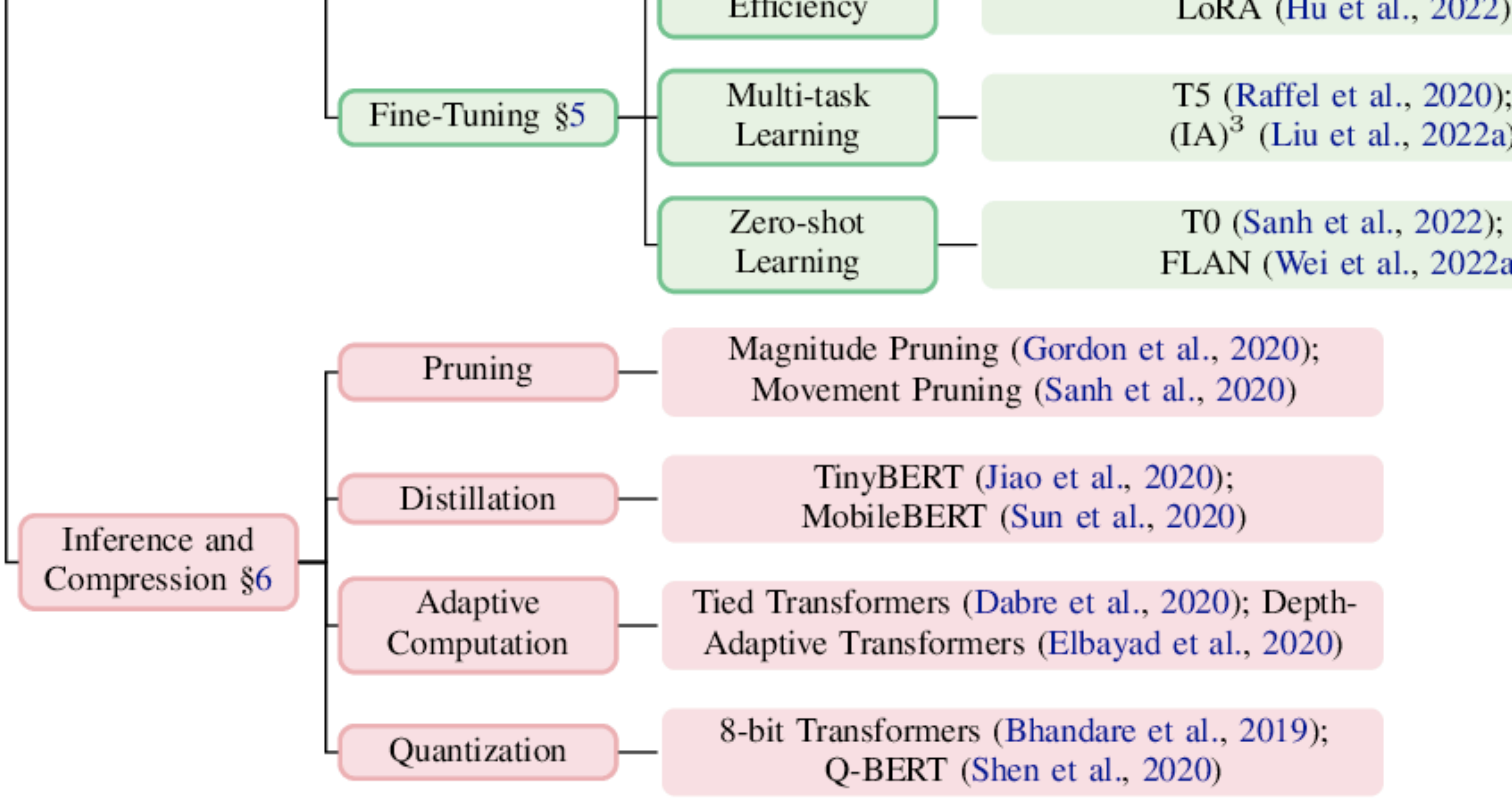
Efficient Methods in NLP




Efficiency in NLP
Retrieval-based
kNN-LM (Khandelwal et al., 2019); RETRO (Borgeaud et al., 2022)
Training
Pre-training §4
Decoder only
GPT-3 (Brown et al., 2020
PaLM (Chowdhery et al., 20
Encoder only
BERT (Devlin et al., 2019
ELECTRA (Clark et al., 202
Encoder-Decoder
T5 (Raffel et al., 2020);
BART (Lewis et al., 2020a
Fine-Tuning §5
Parameter-Efficiency
Adapters (Houlsby et al., 201
LoRA (Hu et al., 2022)
Multi-task Learning
T5 (Raffel et al., 2020);
(IA)3 (Liu et al., 2022a)
Zero-shot Learning
T0 (Sanh et al., 2022);
FLAN (Wei et al., 2022a)
Inference and Compression §6
Pruning
Magnitude Pruning (Gordon et al., 2020); Movement Pruning (Sanh et al., 2020)
Distillation
TinyBERT (Jiao et al., 2020); MobileBERT (Sun et al., 2020)
Adaptive Computation
Tied Transformers (Dabre et al., 2020); Depth-Adaptive Transformers (Elbayad et al., 2020)

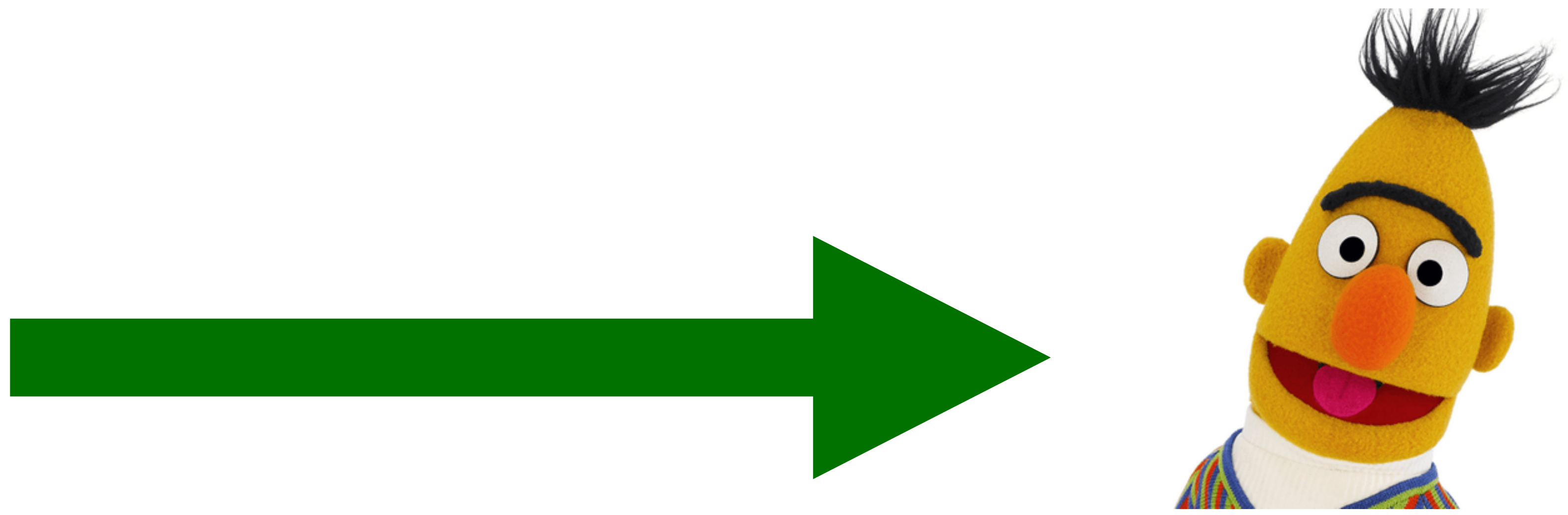
Efficient Methods in NLP




Efficiency
LoRA (Hu et al., 2022)
Fine-Tuning §5
Multi-task Learning
T5 (Raffel et al., 2020);
(IA)3 (Liu et al., 2022a
Zero-shot Learning
T0 (Sanh et al., 2022);
FLAN (Wei et al., 2022a
Inference and Compression §6
Pruning
Magnitude Pruning (Gordon et al., 2020); Movement Pruning (Sanh et al., 2020)
Distillation
TinyBERT (Jiao et al., 2020); MobileBERT (Sun et al., 2020)
Adaptive Computation
Tied Transformers (Dabre et al., 2020); Depth-Adaptive Transformers (Elbayad et al., 2020)
Quantization
8-bit Transformers (Bhandare et al., 2019); Q-BERT (Shen et al., 2020)

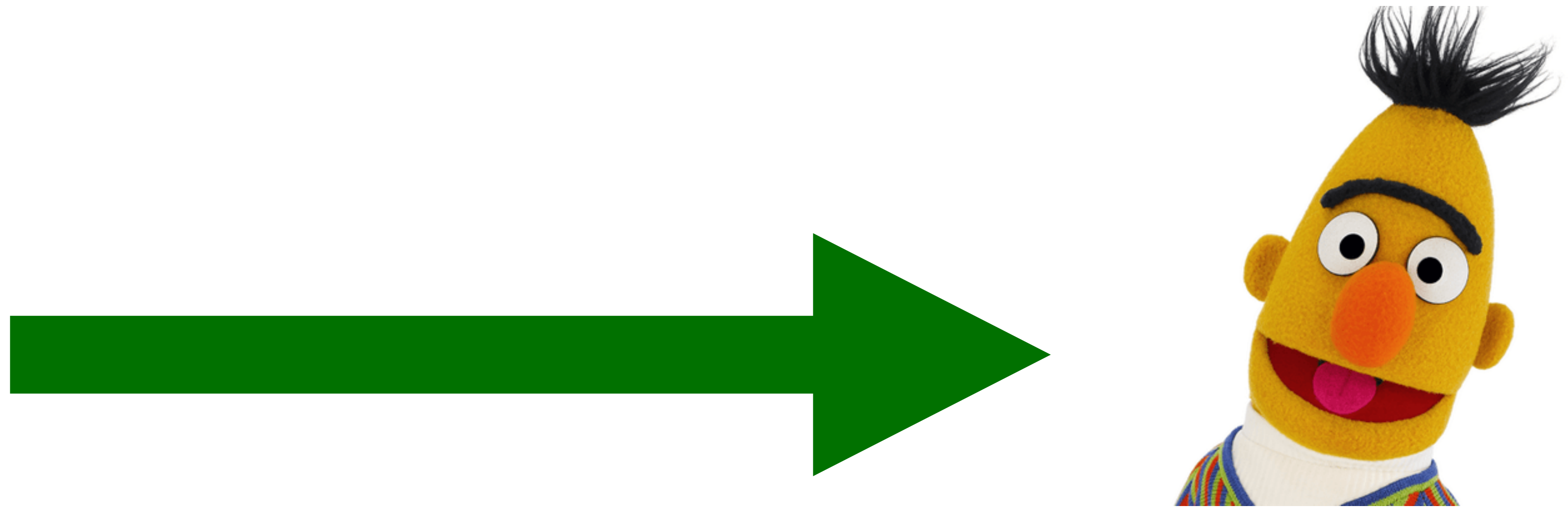
Efficient Inference

Efficient Inference



- **Model distillation**
  - Aka, student/teacher model
  - Hinton et al., 2015; Sun et al., 2019; Sanh et al., 2019

Efficient Inference



- **Model distillation**
  - Aka, student/teacher model
  - Hinton et al., 2015; Sun et al., 2019; Sanh et al., 2019
- **Pruning / Structural Pruning**
  - Han et al., 2016;Lee et al., 2019; Frankle & Corbin, 2019; Gordon et al., 2018; Michel et al., 2019; Fan et al., 2020
  - Dodge, **S.**, et al., 2019

Efficient Inference

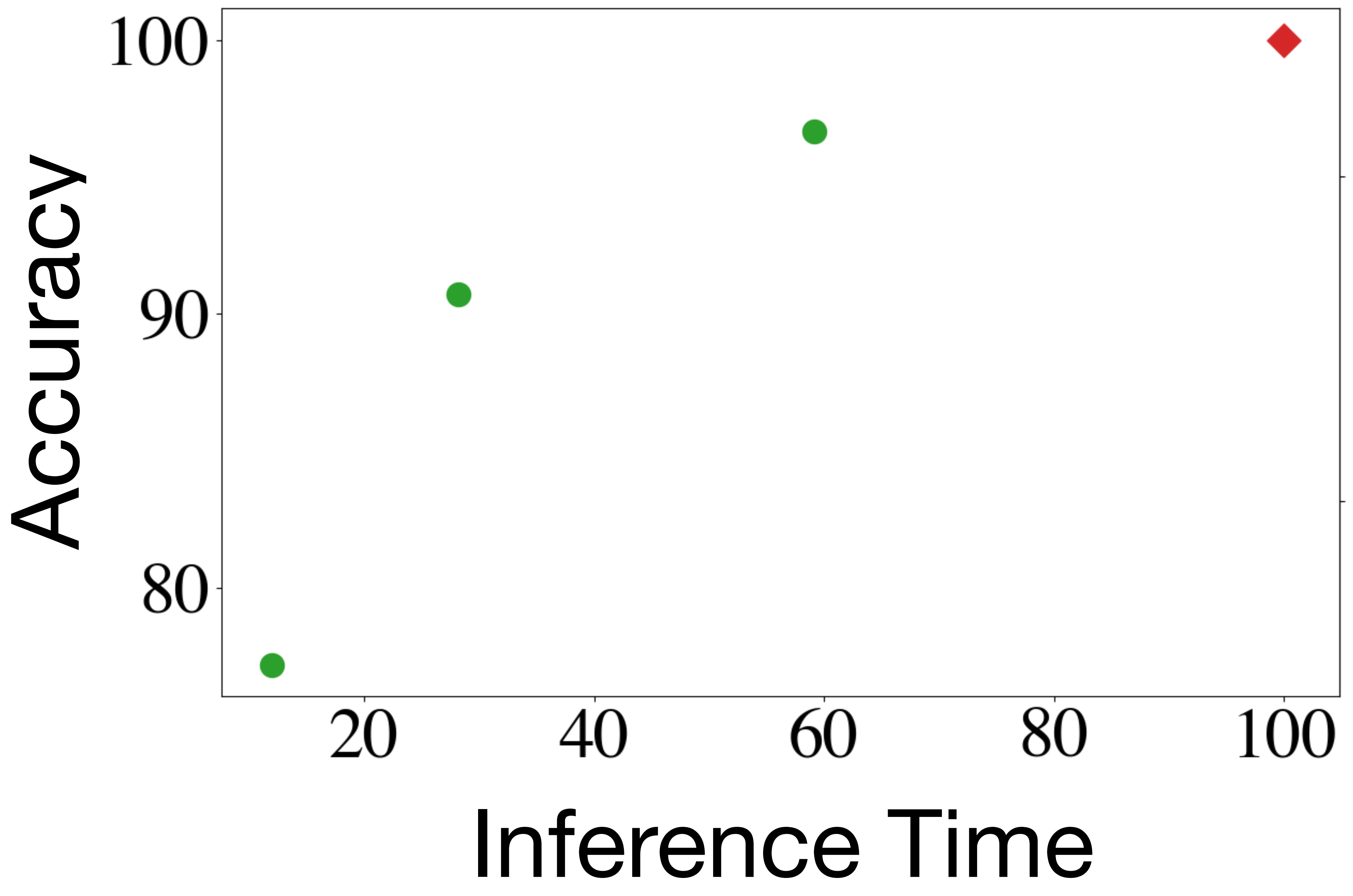
- **Model distillation**
  - Aka, student/teacher model
  - Hinton et al., 2015; Sun et al., 2019; Sanh et al., 2019
- **Pruning / Structural Pruning**
  - Han et al., 2016;Lee et al., 2019; Frankle & Corbin, 2019; Gordon et al., 2018; Michel et al., 2019; Fan et al., 2020
  - Dodge, **S.**, et al., 2019
- **Quantization**
  - Gong et al., 2014; Zafrir et al., 2019; Shen et al., 2019

Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

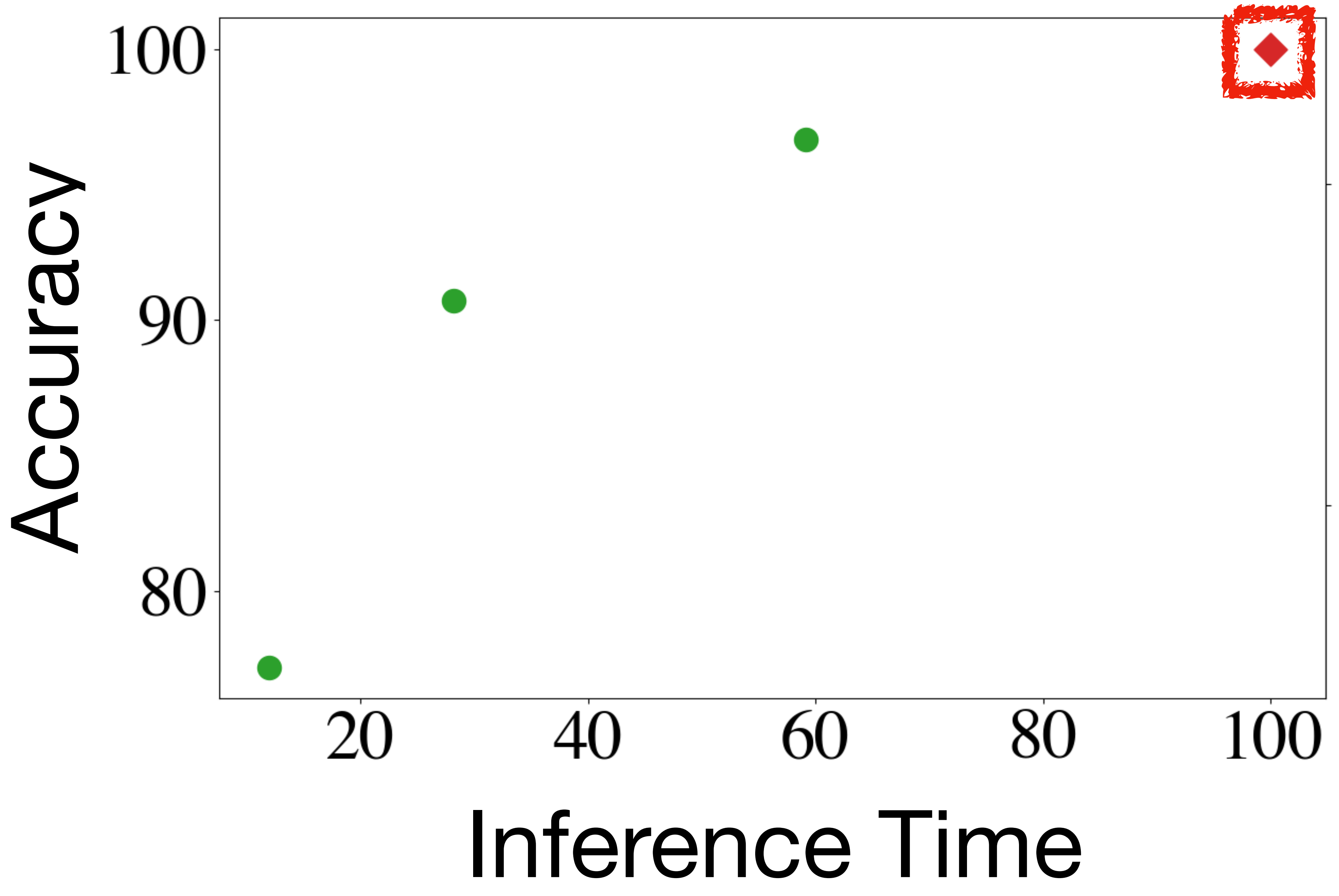


Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

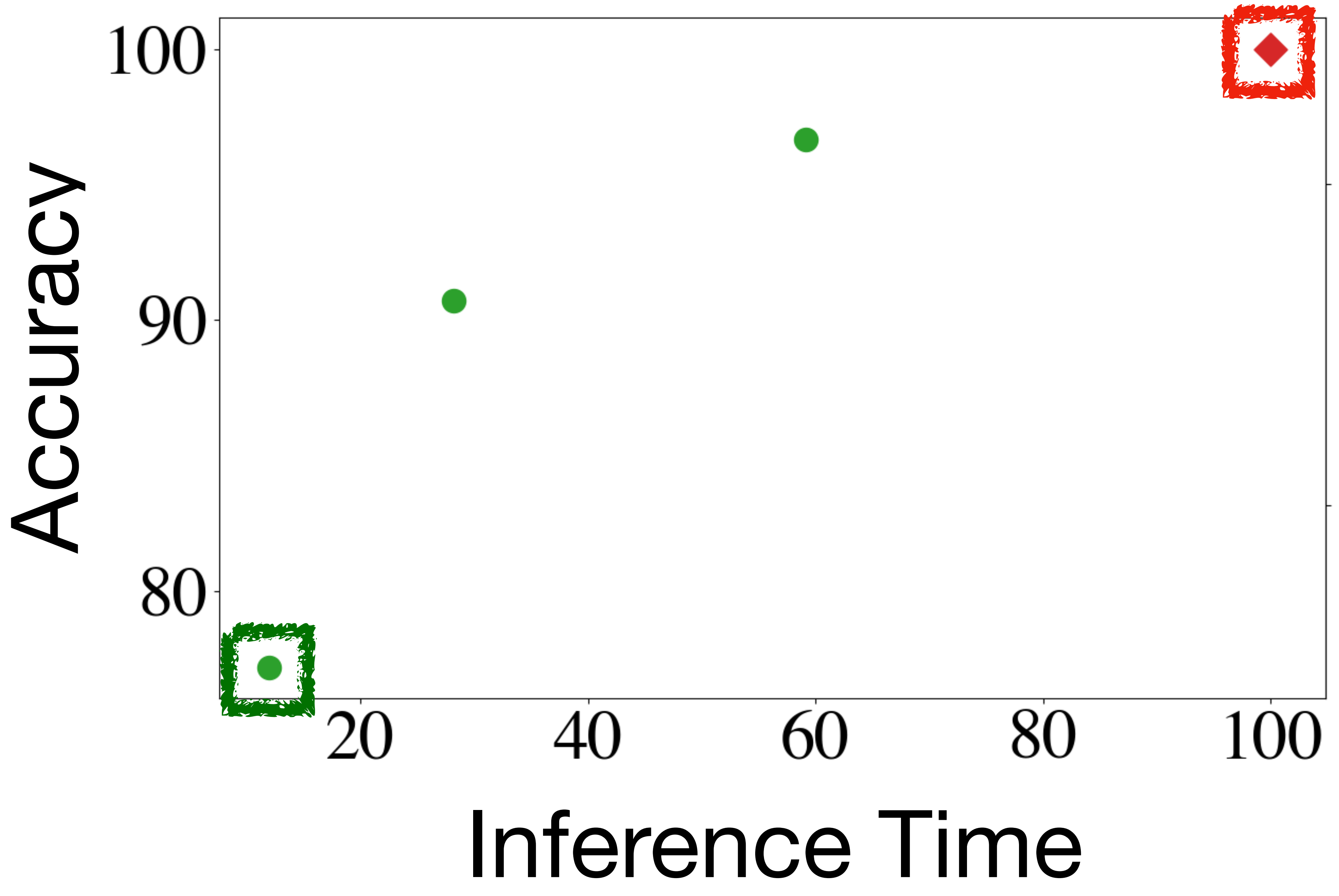


Matching Model and Instance Complexity

S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

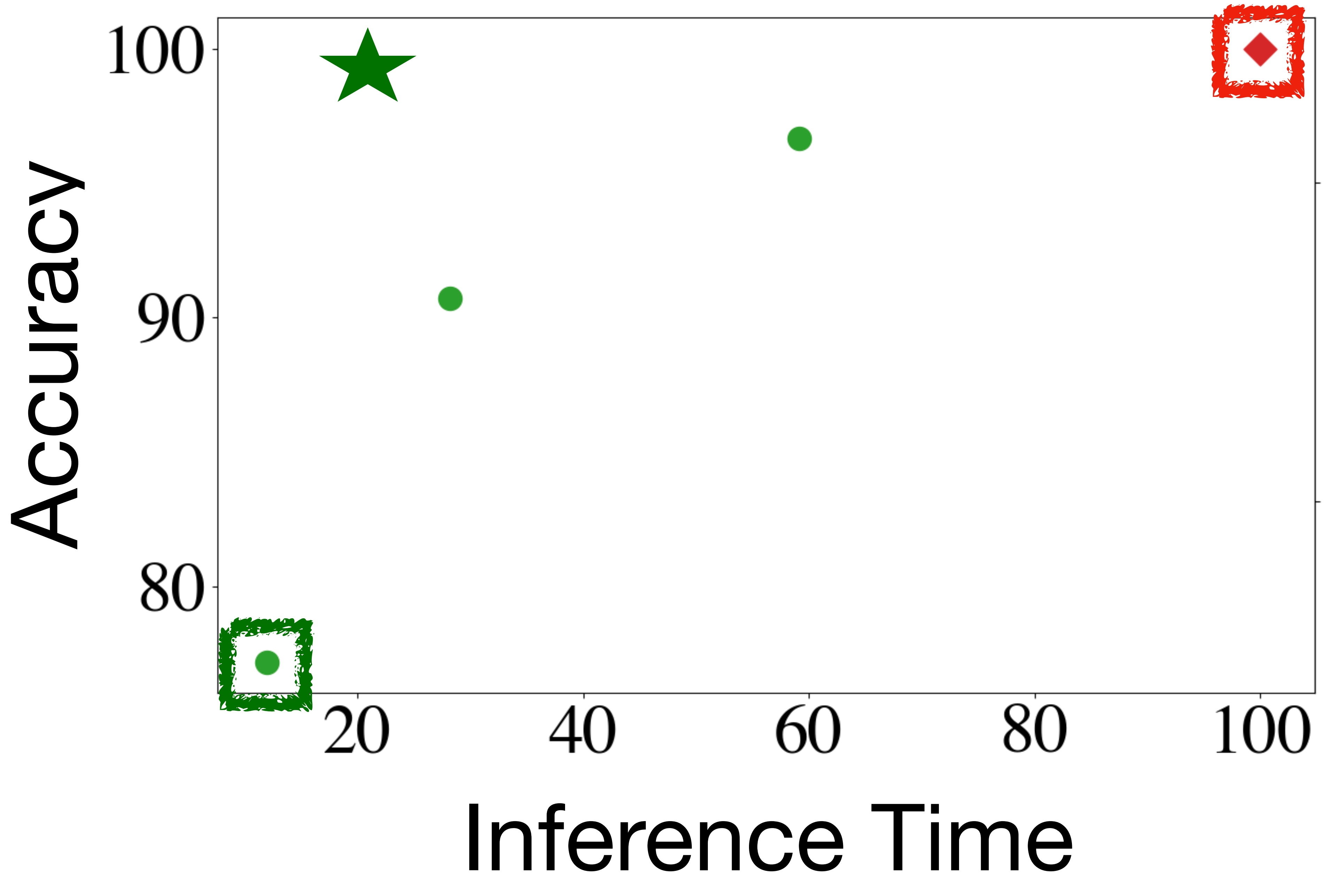


Matching Model and Instance Complexity

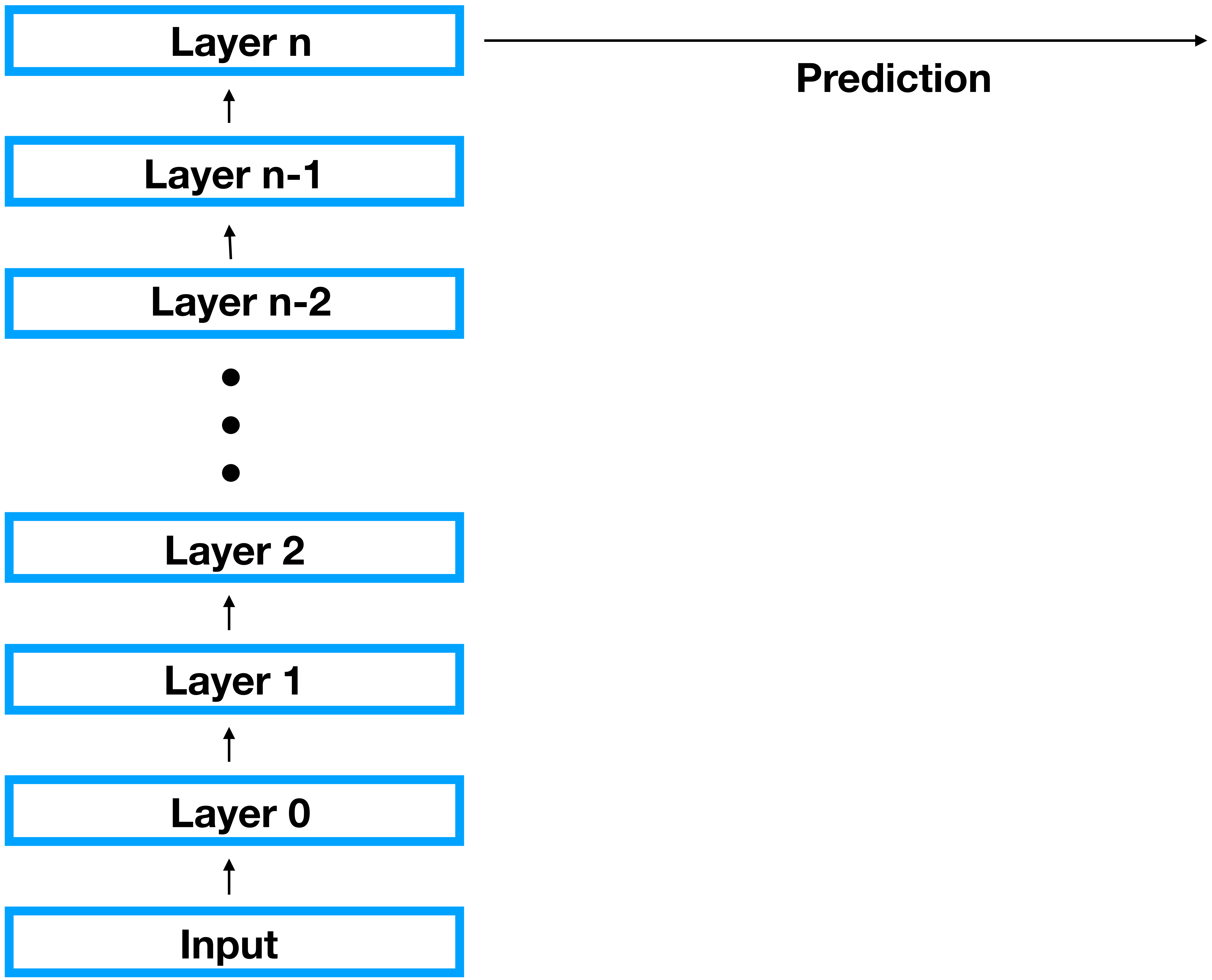
S. et al., ACL 2020

Run an efficient model on "easy" instances,

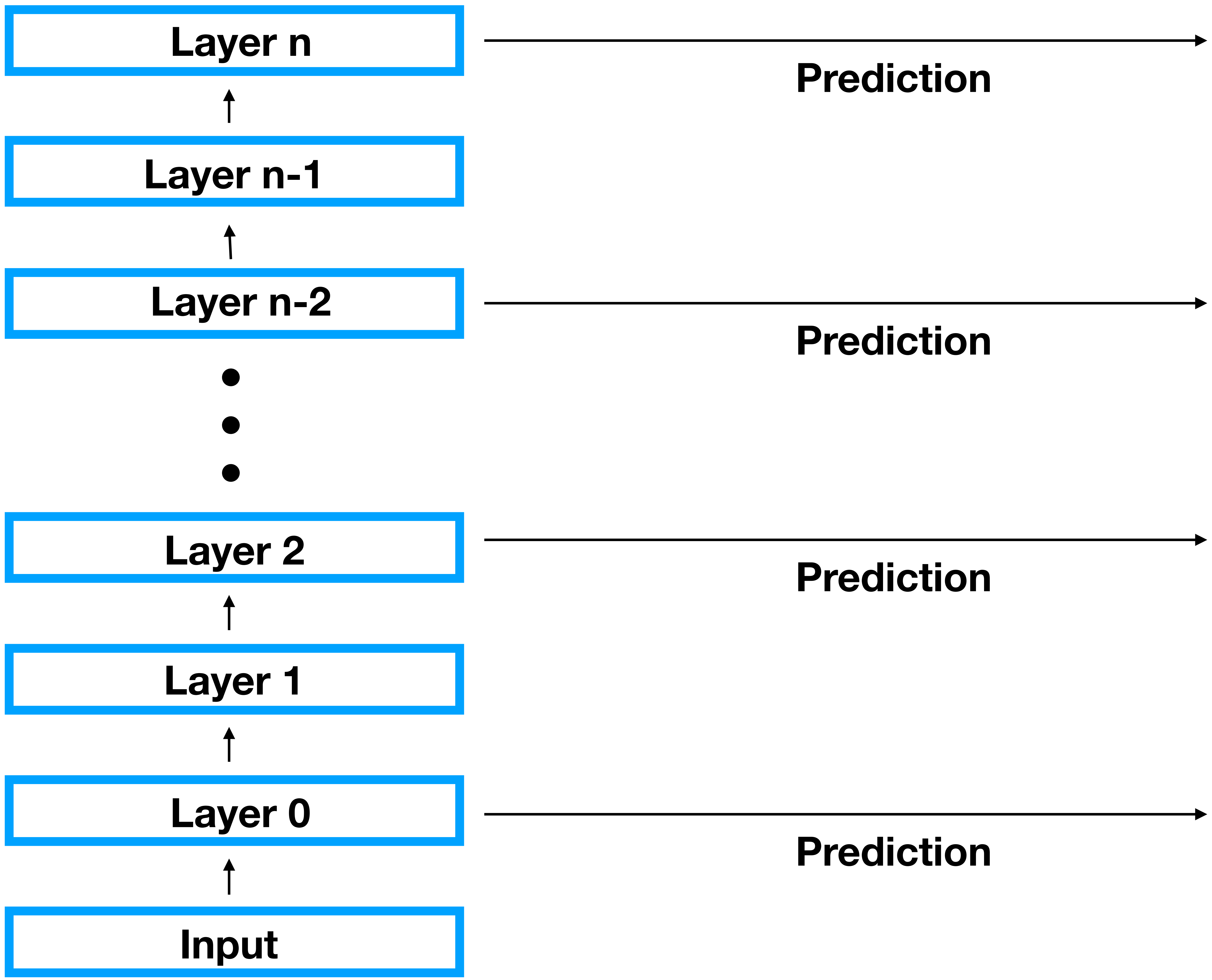
and an expensive model on "hard" instances



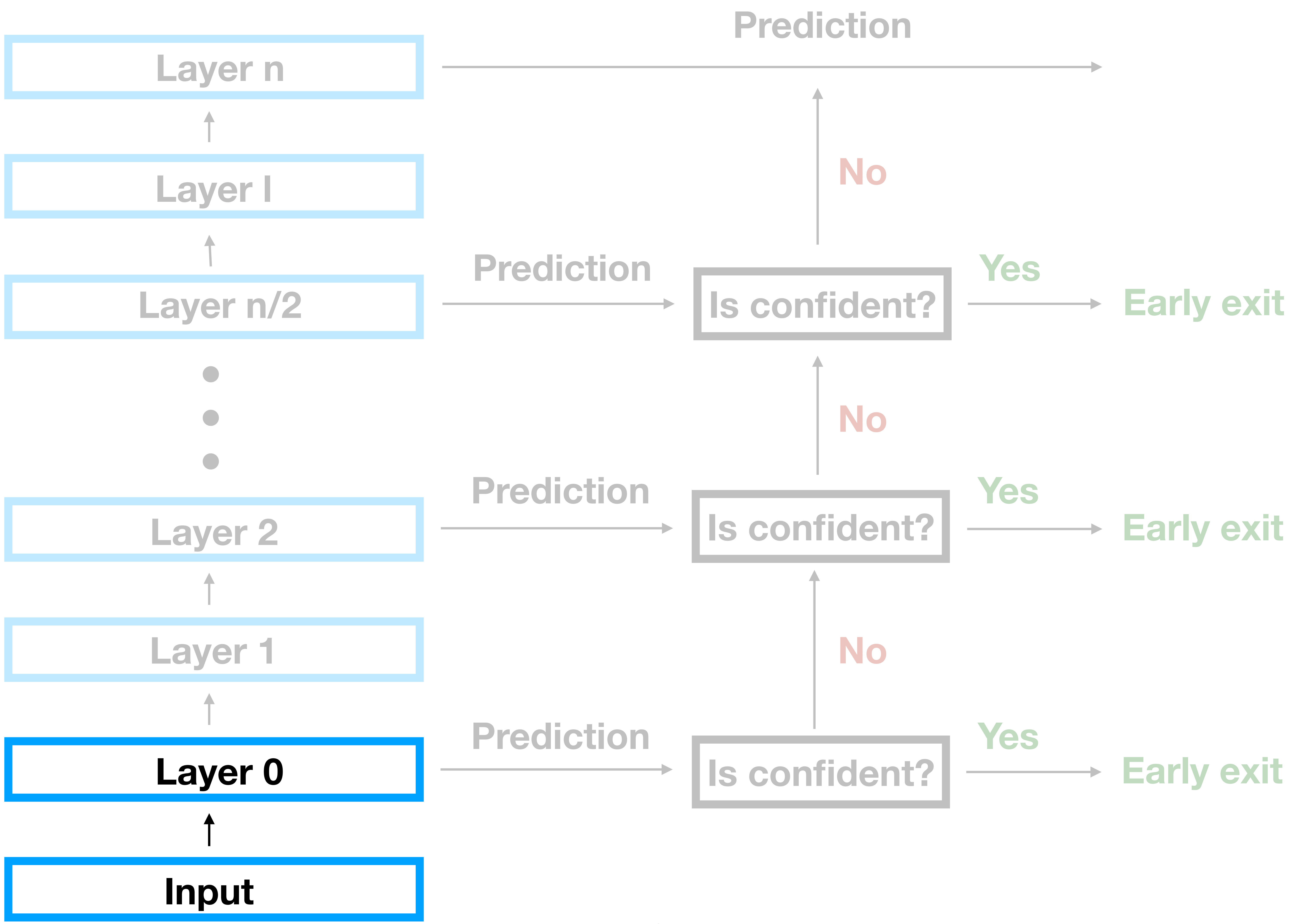
Our Approach: **Training Time**



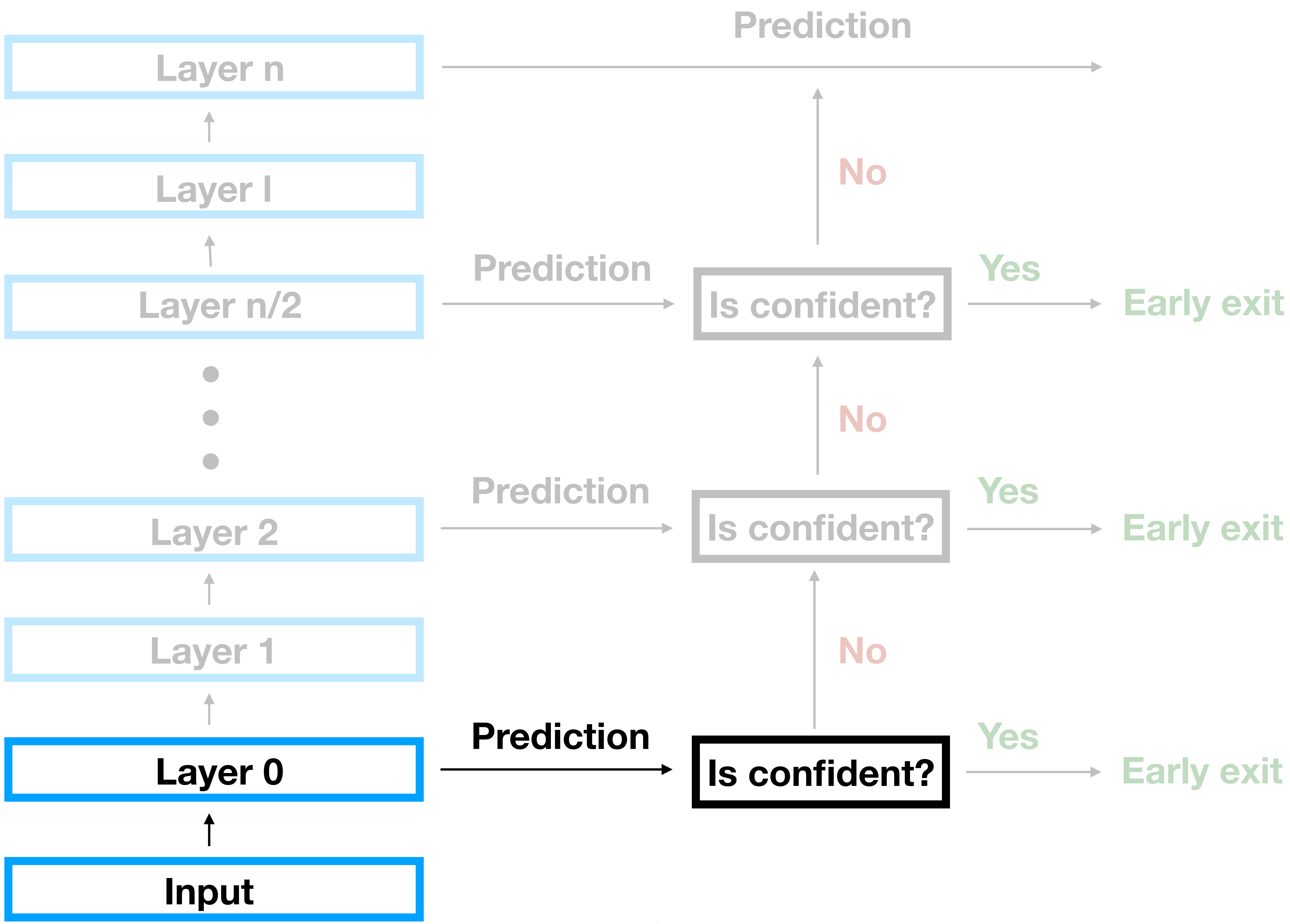
Our Approach: **Training Time**



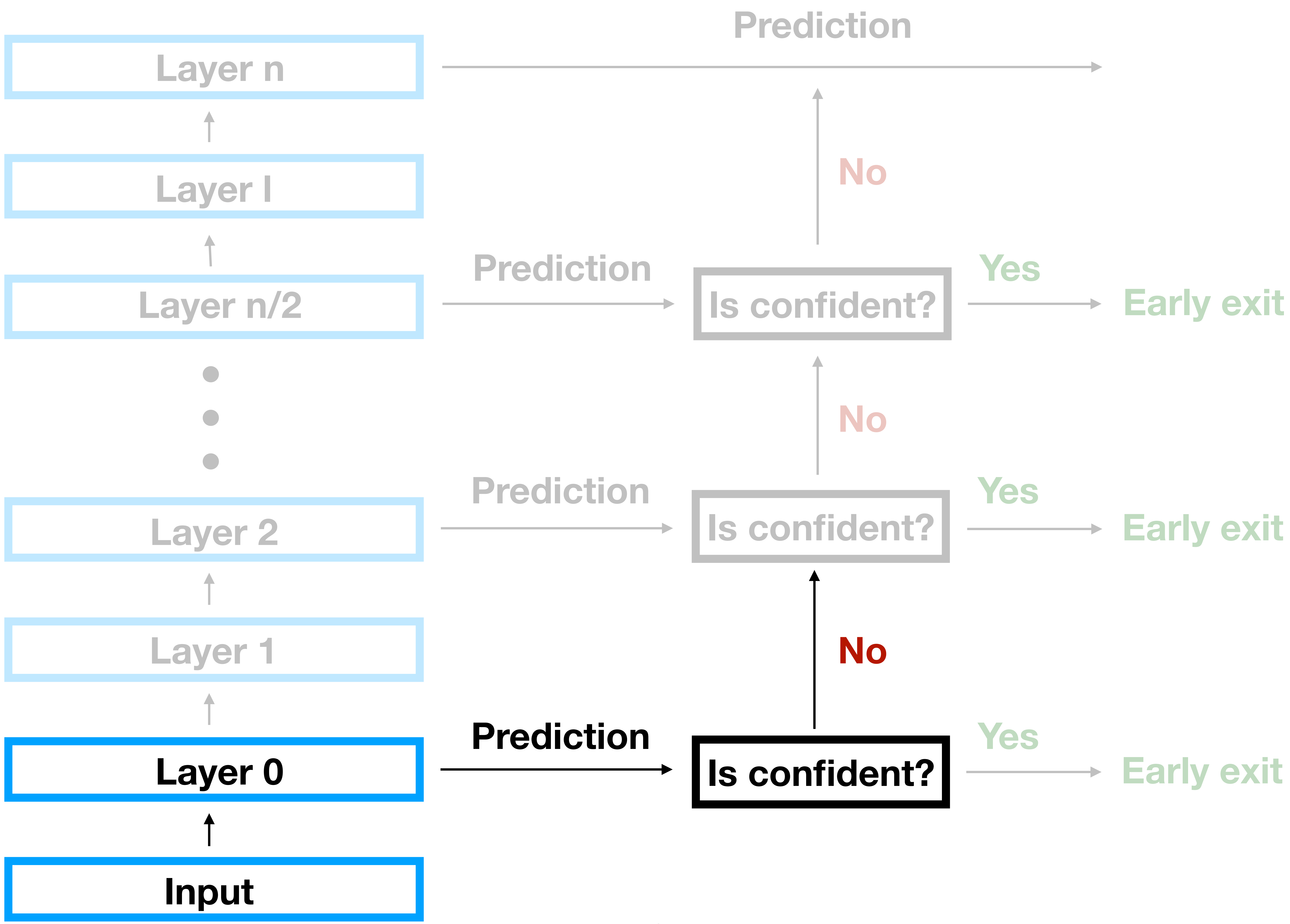
Our Approach: Test Time



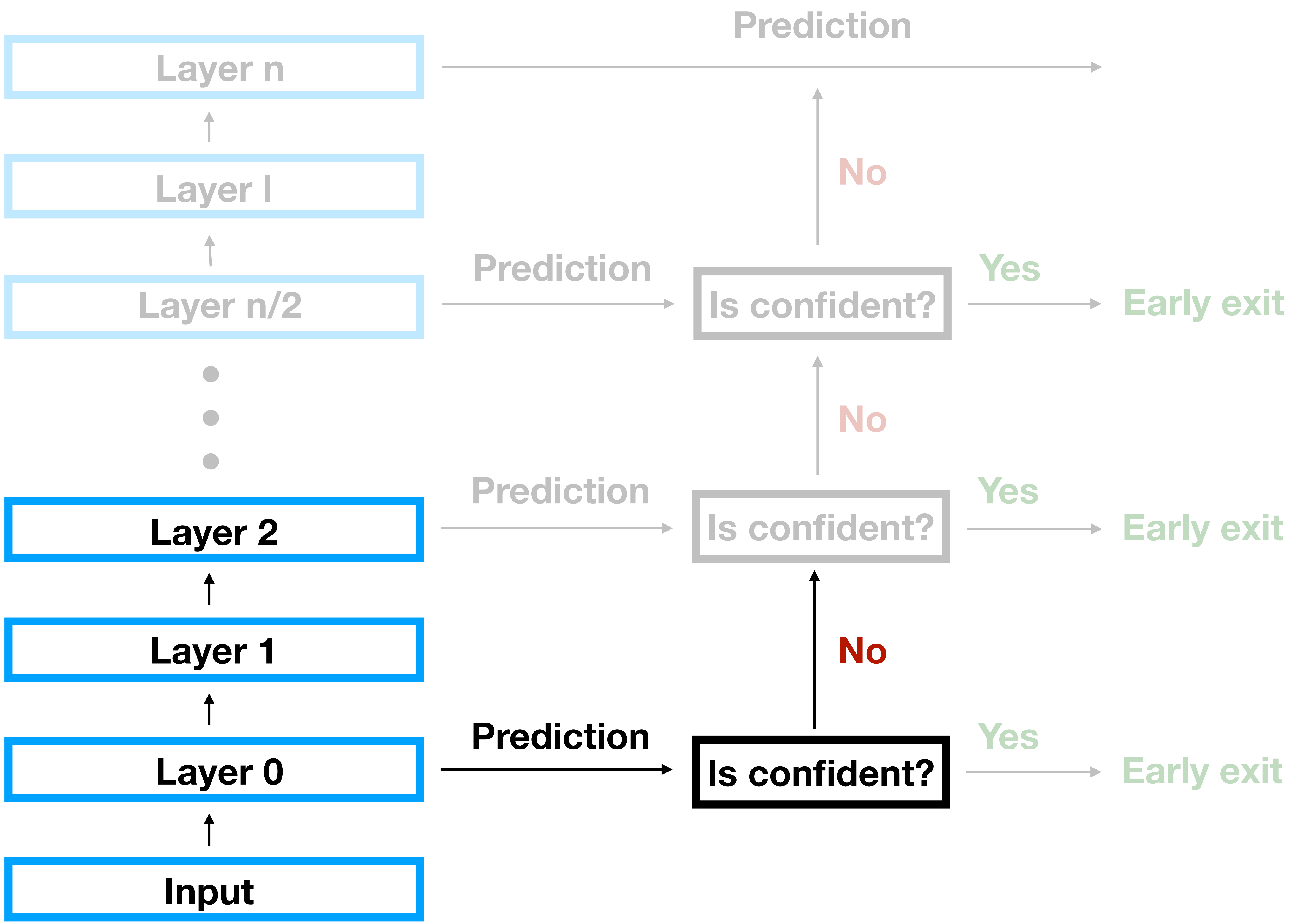
Our Approach: Test Time



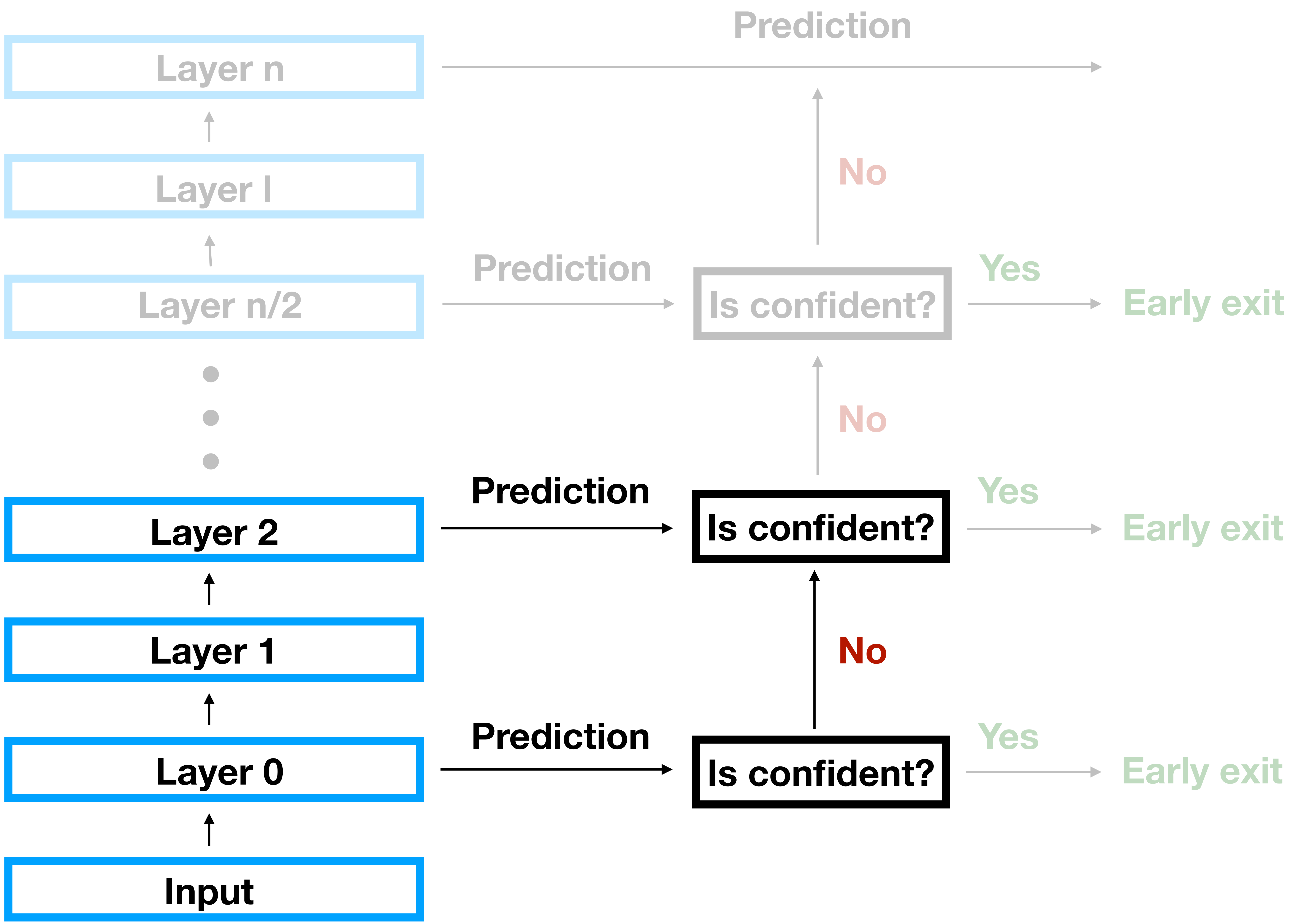
Our Approach: Test Time



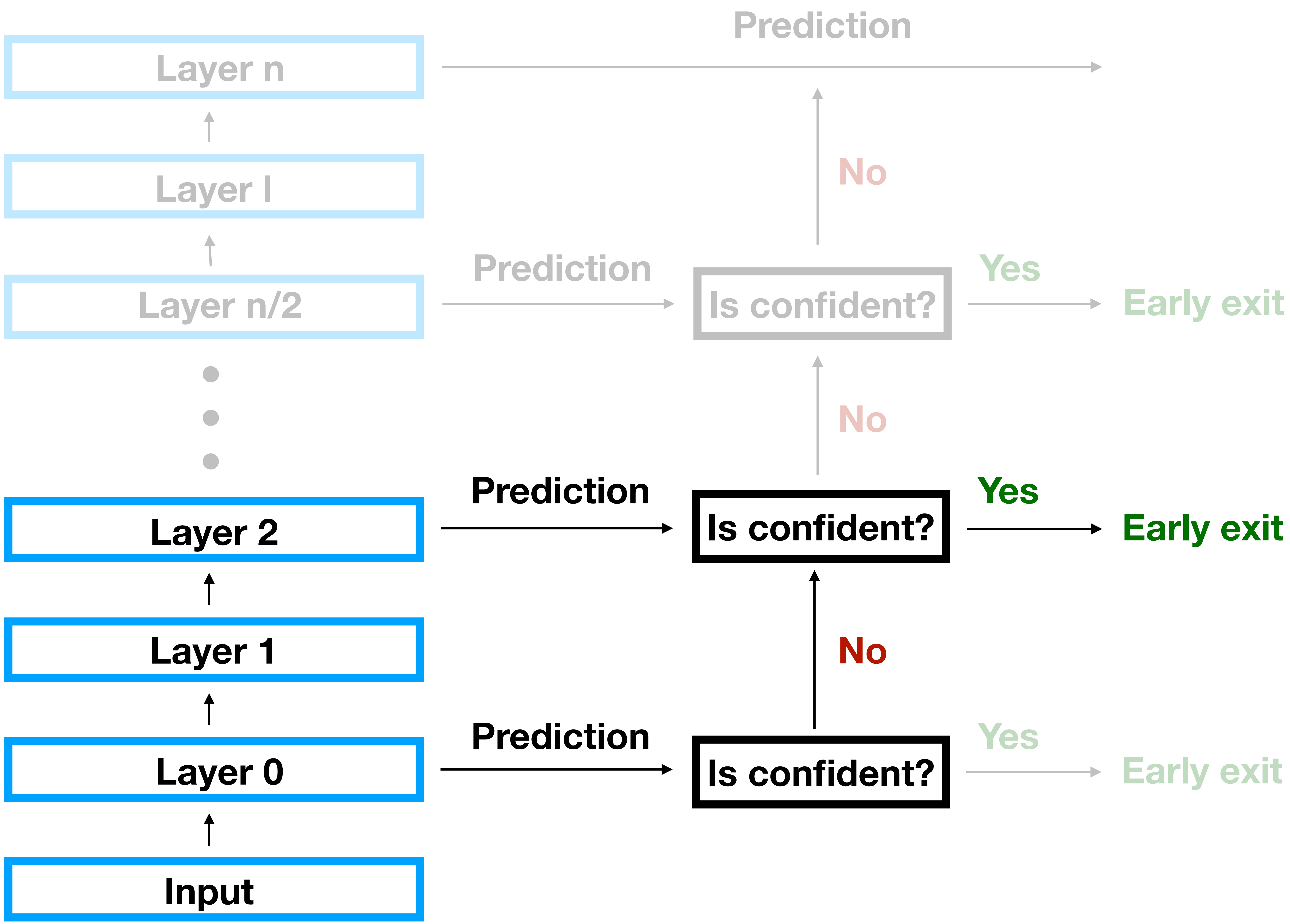
Our Approach: Test Time



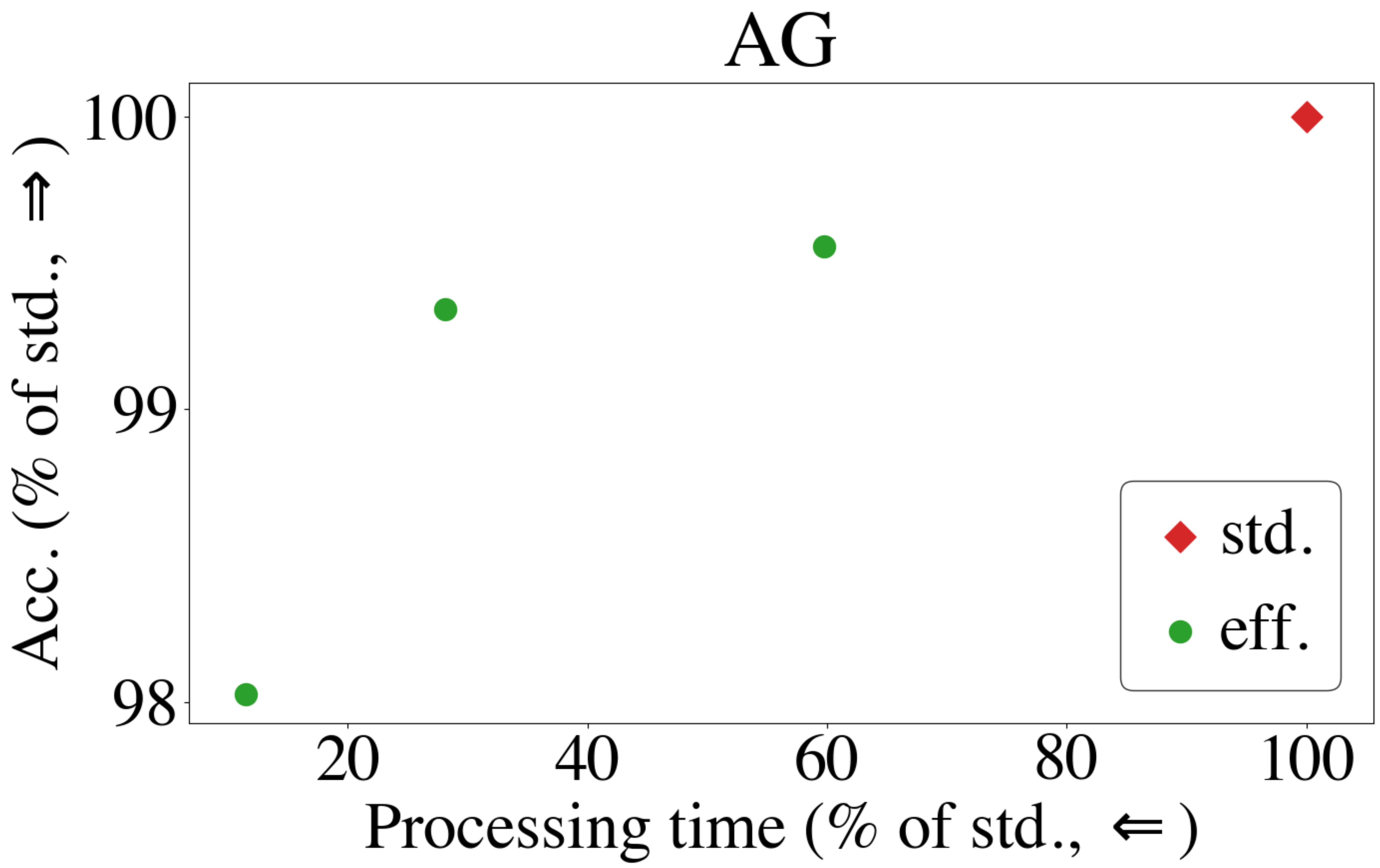
Our Approach: Test Time



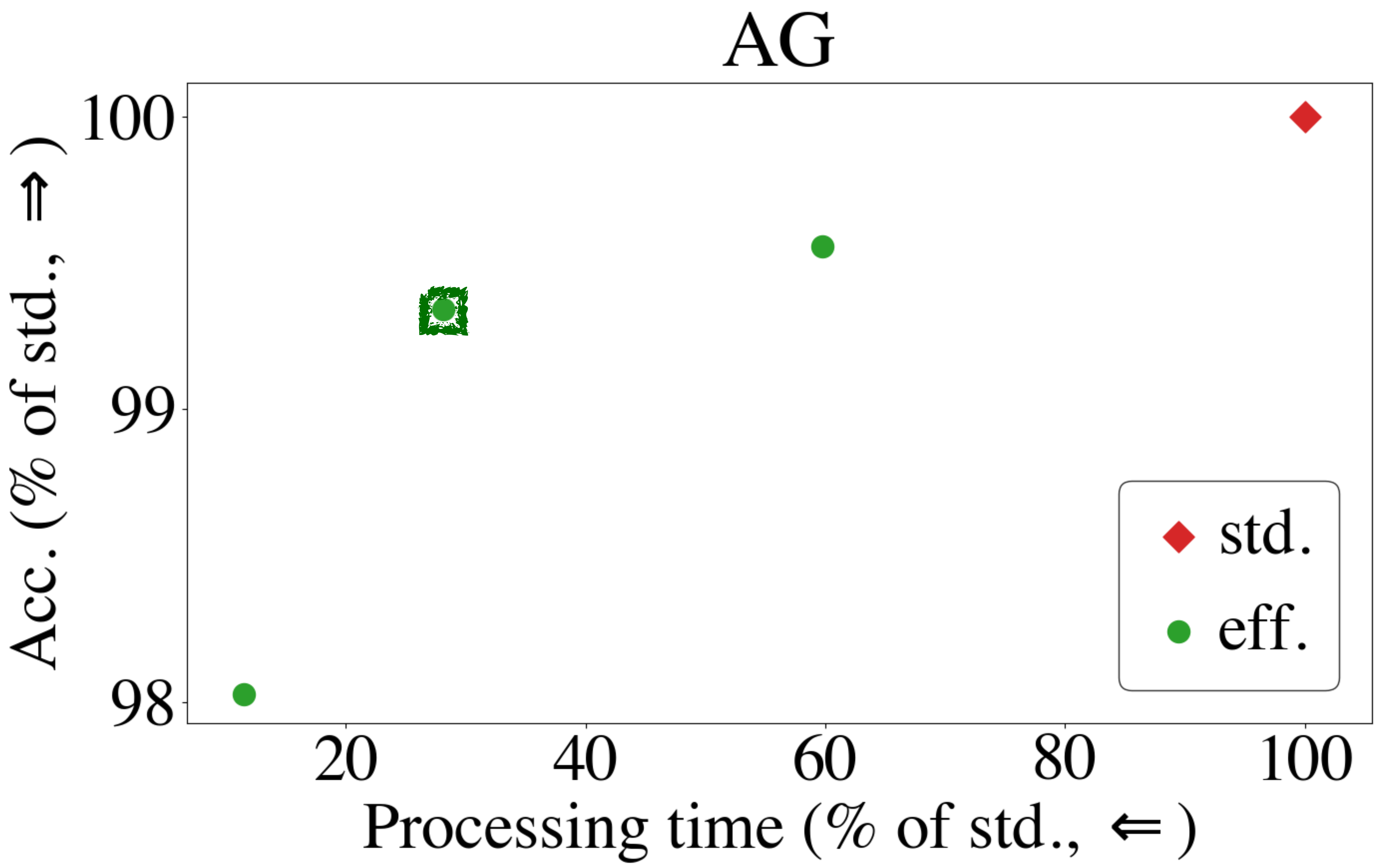
Our Approach: Test Time



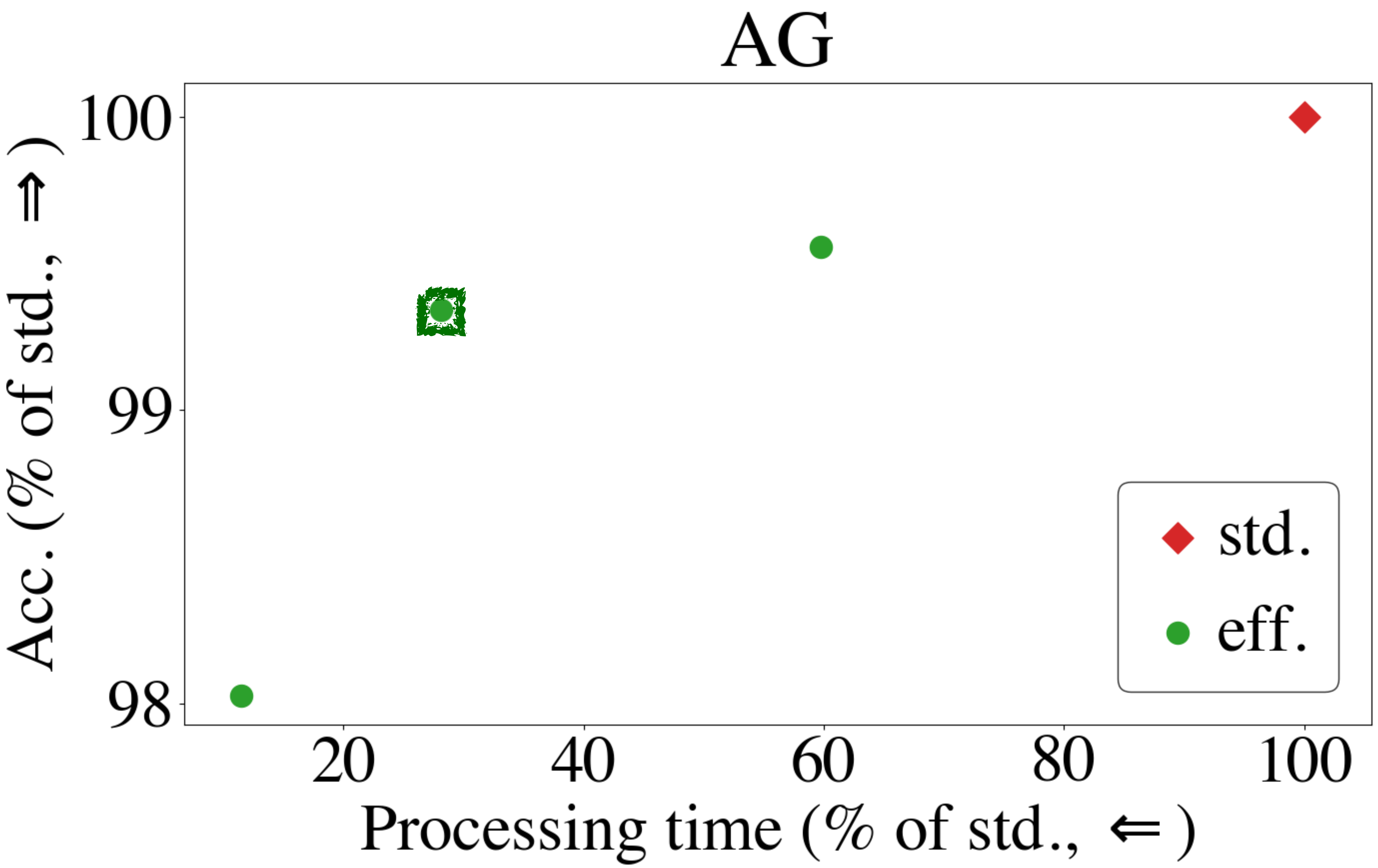
Strong Baselines!



Strong Baselines!

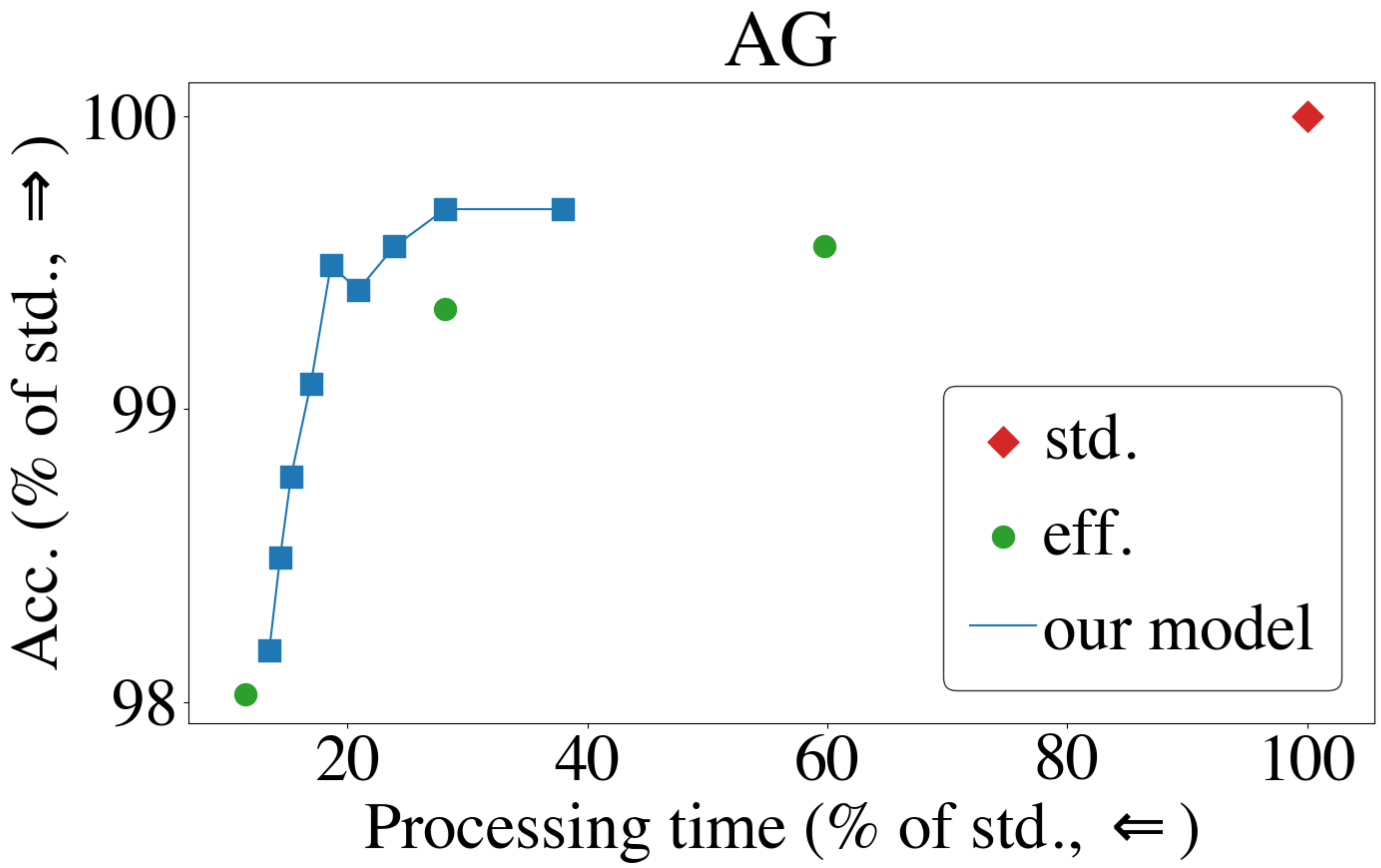


Strong Baselines!

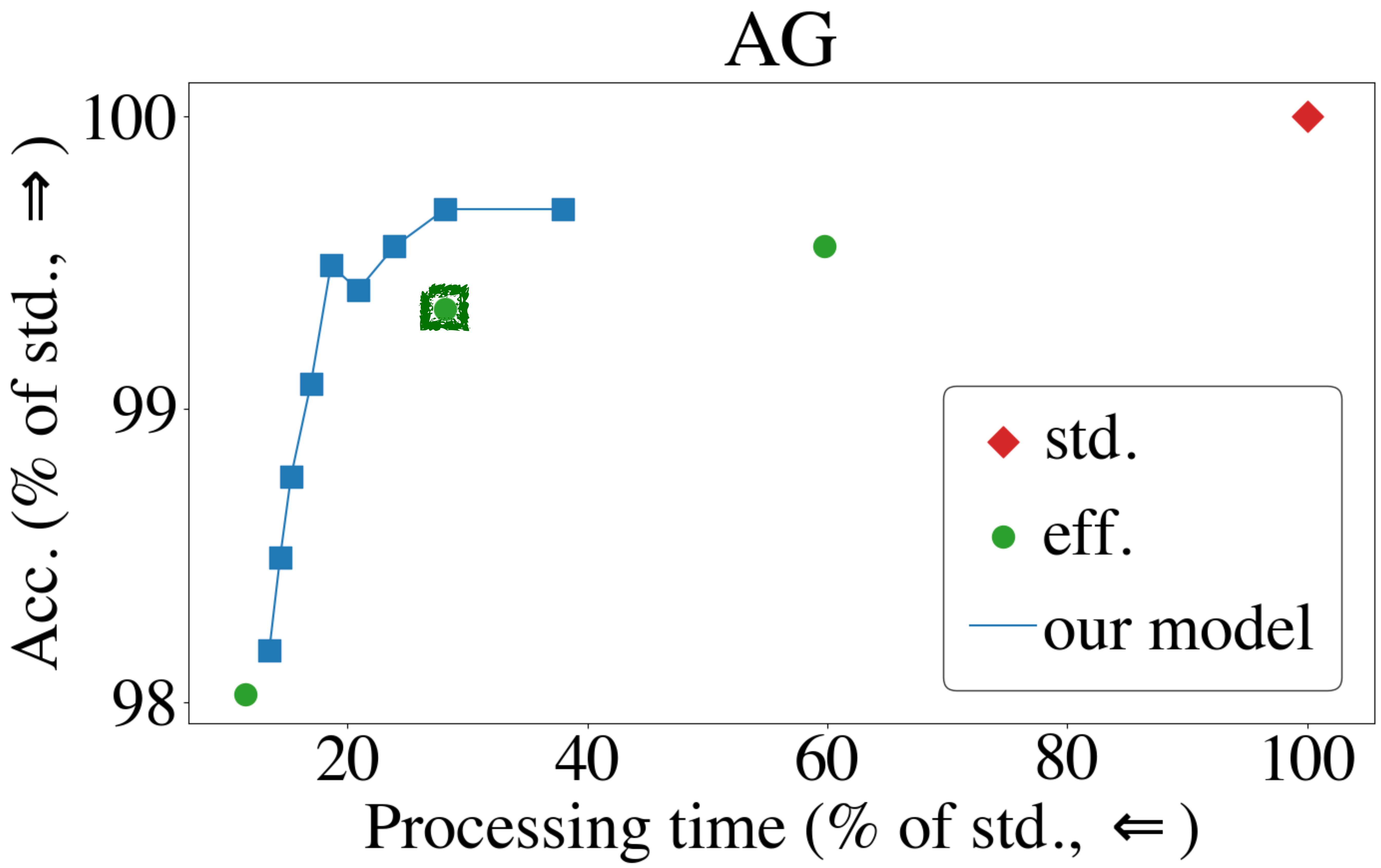


3 times faster, within 1% of full model

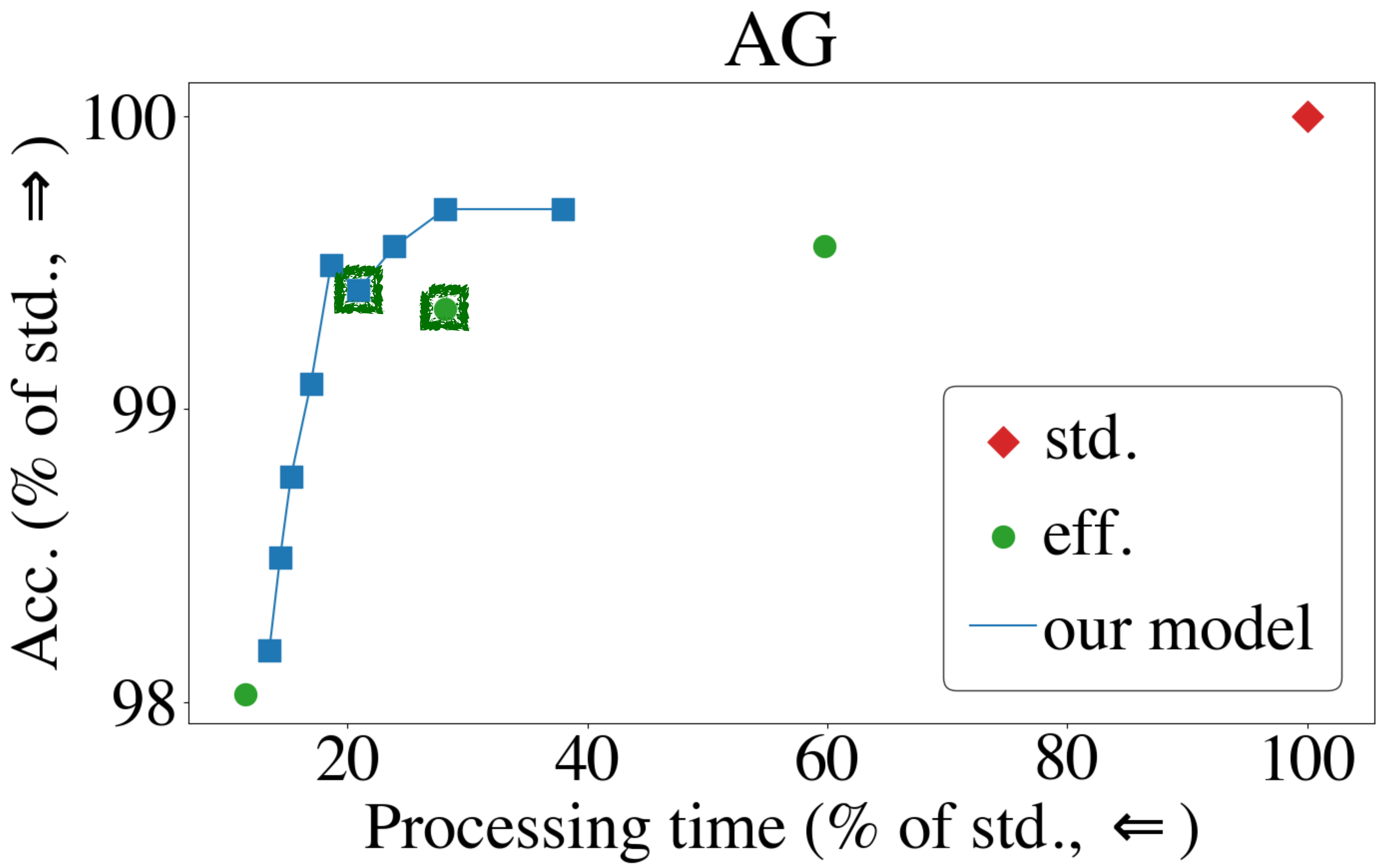
Better Speed/Accuracy Tradeoff



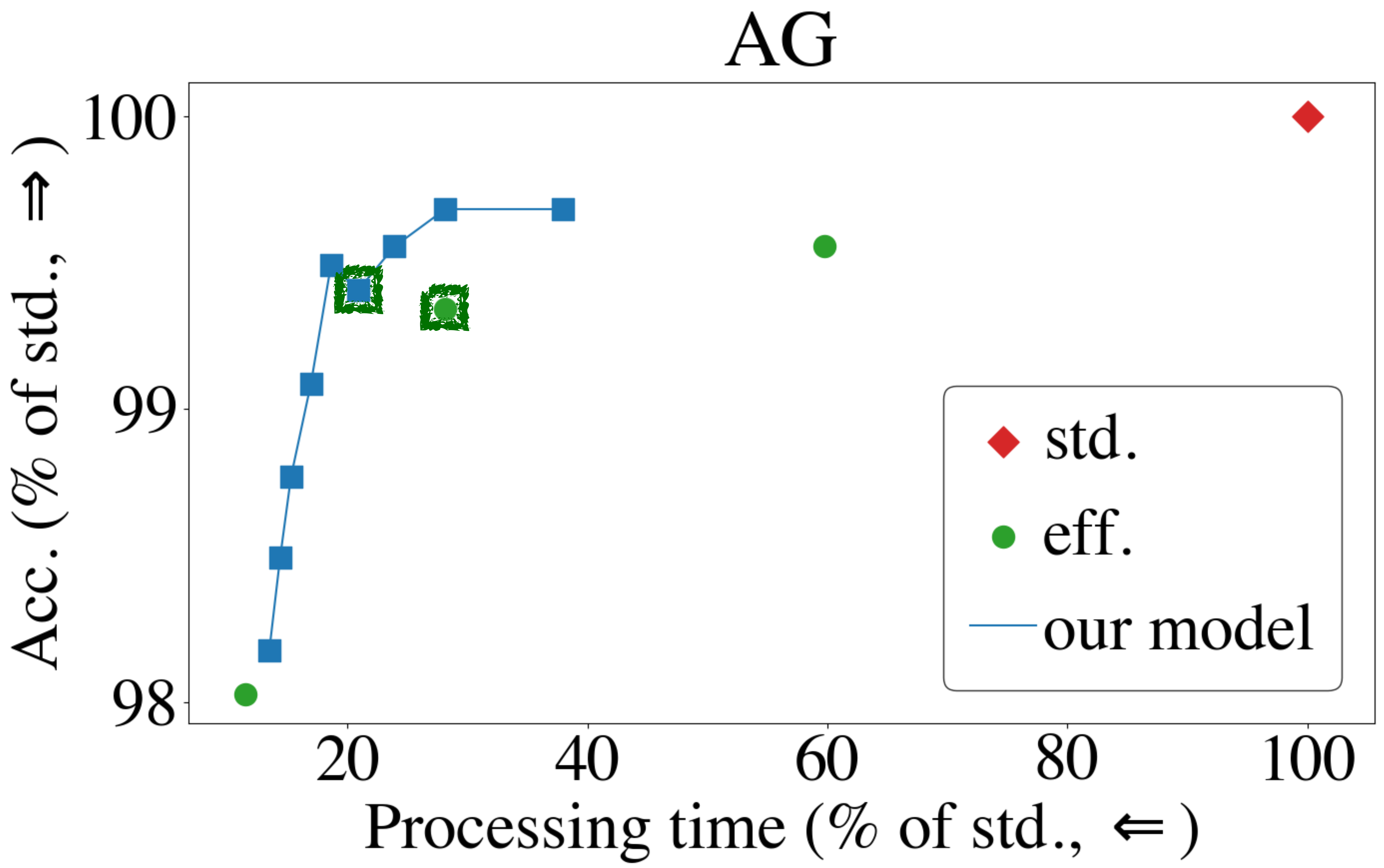
Better Speed/Accuracy Tradeoff



Better Speed/Accuracy Tradeoff



Better Speed/Accuracy Tradeoff



5 times faster, within 1% of full model

Efficiency

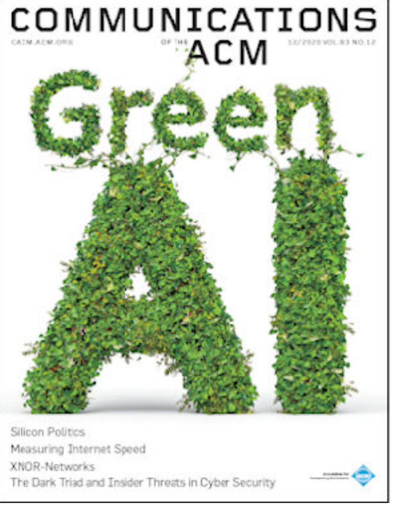
Open Questions

- What makes a good sparse structure?

Efficiency

Open Questions

- What makes a good sparse structure?
- Combining different methods

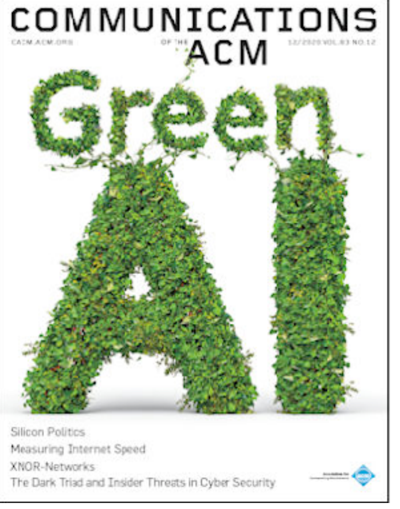


Think Green!

- Red AI
  - Problems: inclusiveness, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results

  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy



Think Green!

- Red AI
  - Problems: inclusiveness, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results

  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy