Resolving Discrepancies in Compute-Optimal Scaling of Language Models

Tomer Porian

Mitchell Wortsman

Jenia Jitsev

Ludwig Schmidt

Yair Carmon

TEL AVIV UNIVERSITY

UNIVERSITY of WASHINGTON

JÜLICH FORSCHUNGSZENTRUM

Compute-optimal scaling laws

Compute-optimal scaling laws

Given a fixed compute budget C, what is the optimal model size $N^\star(C)$?

Compute-optimal scaling laws

Given a fixed compute budget C, what is the optimal model size $N^\star(C)$?

$$N^\star(C) \propto C^a$$

Discrepancy in scaling laws

Discrepancy in scaling laws

Kaplan et al.: $N^\star(C) \propto C^{0.88}$

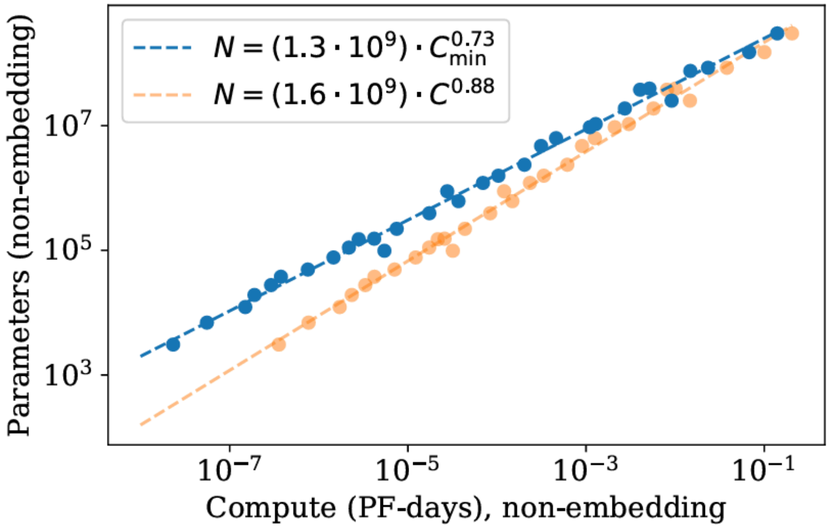
(After post-processing: $N^\star(C) \propto C^{0.73}$)

J. Kaplan et al.
Scaling laws for neural language models.
arXiv:2001.08361, 2020.

Discrepancy in scaling laws

Kaplan et al.: $N^\star(C) \propto C^{0.88}$

(After post-processing: $N^\star(C) \propto C^{0.73}$)



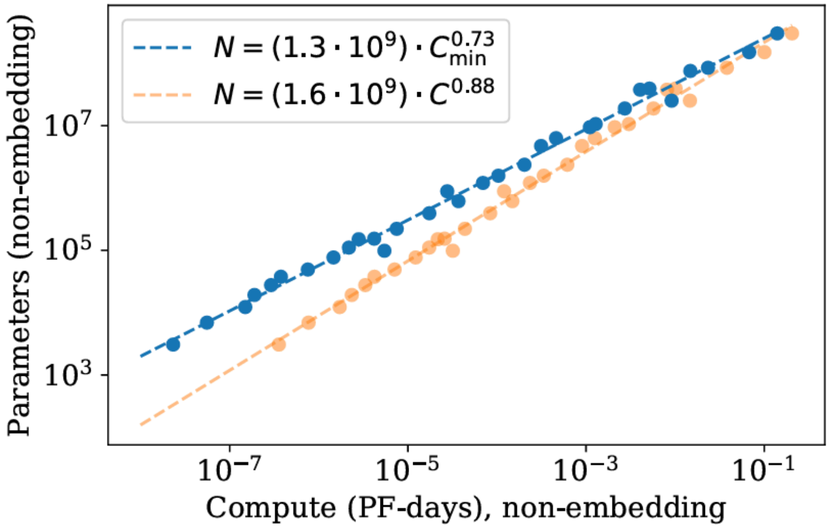
J. Kaplan et al.
Scaling laws for neural language models.
arXiv:2001.08361, 2020.

Discrepancy in scaling laws

Kaplan et al.: $N^\star(C) \propto C^{0.88}$

(After post-processing: $N^\star(C) \propto C^{0.73}$)

Hoffmann et al.: $N^\star(C) \propto C^{0.5}$ (“Chinchilla law”)



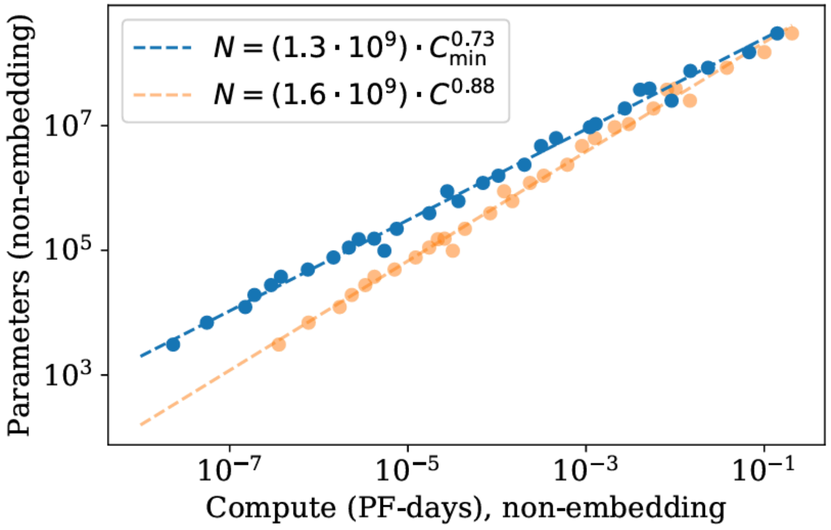
J. Kaplan et al.
Scaling laws for neural language models.
arXiv:2001.08361, 2020.

J. Hoffmann et al.
An empirical analysis of compute-optimal large language model training.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022

Discrepancy in scaling laws

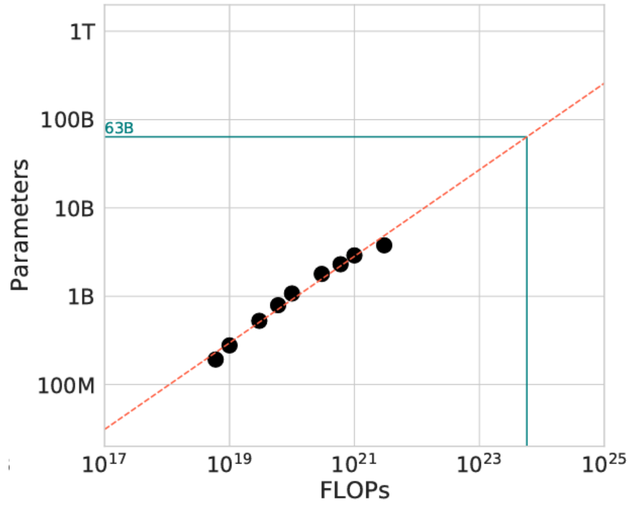
Kaplan et al.: $N^\star(C) \propto C^{0.88}$

(After post-processing: $N^\star(C) \propto C^{0.73}$)



J. Kaplan et al.
Scaling laws for neural language models.
arXiv:2001.08361, 2020.

Hoffmann et al.: $N^\star(C) \propto C^{0.5}$ ("Chinchilla law")

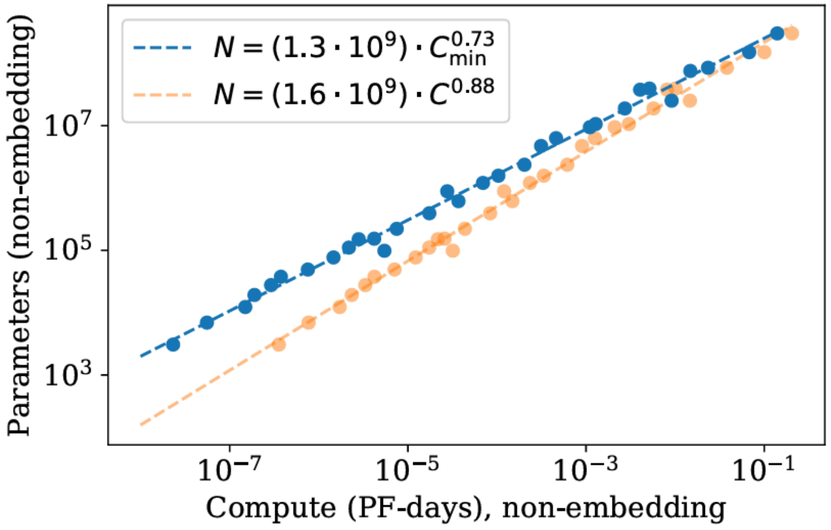


J. Hoffmann et al.
An empirical analysis of compute-optimal large language model training.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022

Discrepancy in scaling laws

Kaplan et al.: $N^\star(C) \propto C^{0.88}$

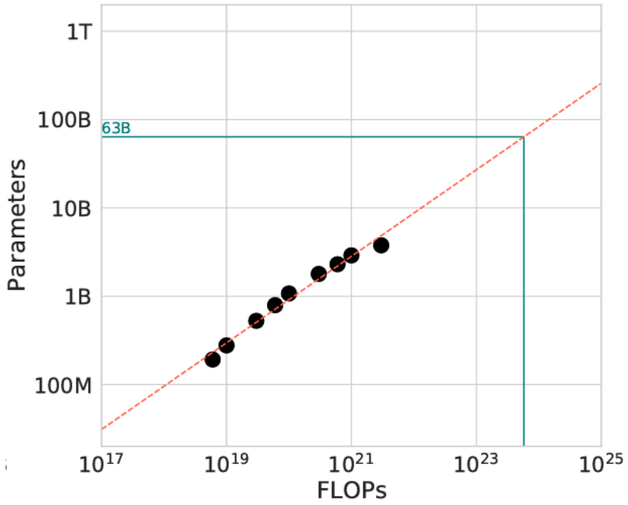
(After post-processing: $N^\star(C) \propto C^{0.73}$)



J. Kaplan et al.
Scaling laws for neural language models.
arXiv:2001.08361, 2020.

Hoffmann et al.: $N^\star(C) \propto C^{0.5}$ ("Chinchilla law")

Conjectured reason: LR decay



J. Hoffmann et al.
An empirical analysis of compute-optimal large language model training.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022

Experimental setup

Experimental setup

- Using the OpenLM library 

Experimental setup

- Using the OpenLM library 

- 16 model architectures, with 5M-901M parameters

Experimental setup

- Using the OpenLM library 

- 16 model architectures, with 5M-901M parameters
- Two open datasets:

Experimental setup

- Using the OpenLM library 

- 16 model architectures, with 5M-901M parameters
- Two open datasets:
  - OpenWebText2 — ~30B tokens, resembles Kaplan et al. dataset

Experimental setup

- Using the OpenLM library 

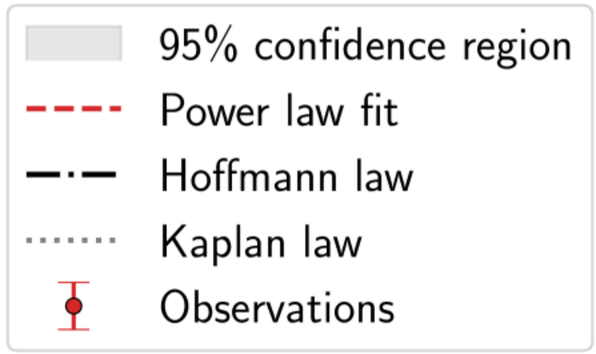
- 16 model architectures, with 5M-901M parameters
- Two open datasets:
  - OpenWebText2 — ~30B tokens, resembles Kaplan et al. dataset
  - RefinedWeb— ~600B tokens from CommonCrawl, resembles half of the data mix in Hoffmann et al.

Three reasons

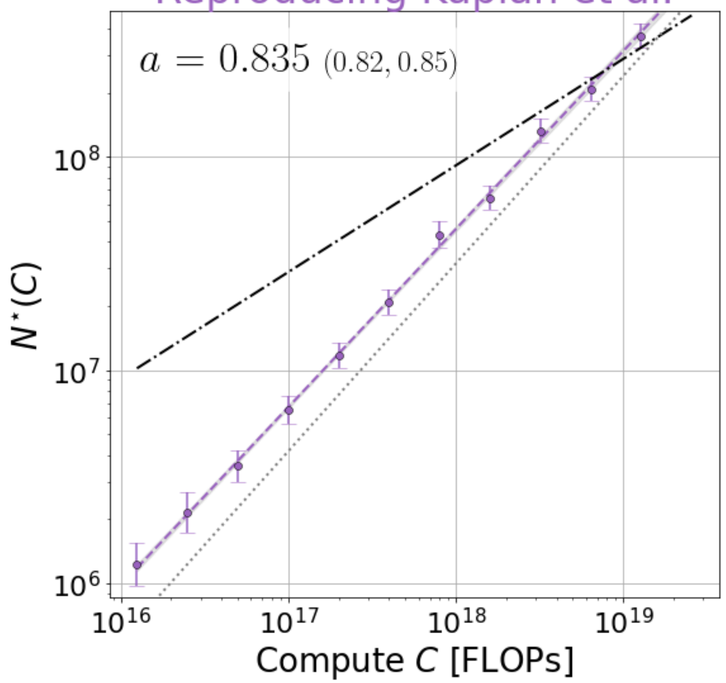
for the discrepancy

Three reasons

for the discrepancy

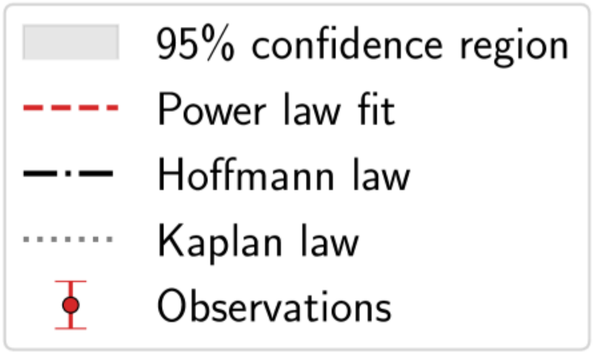


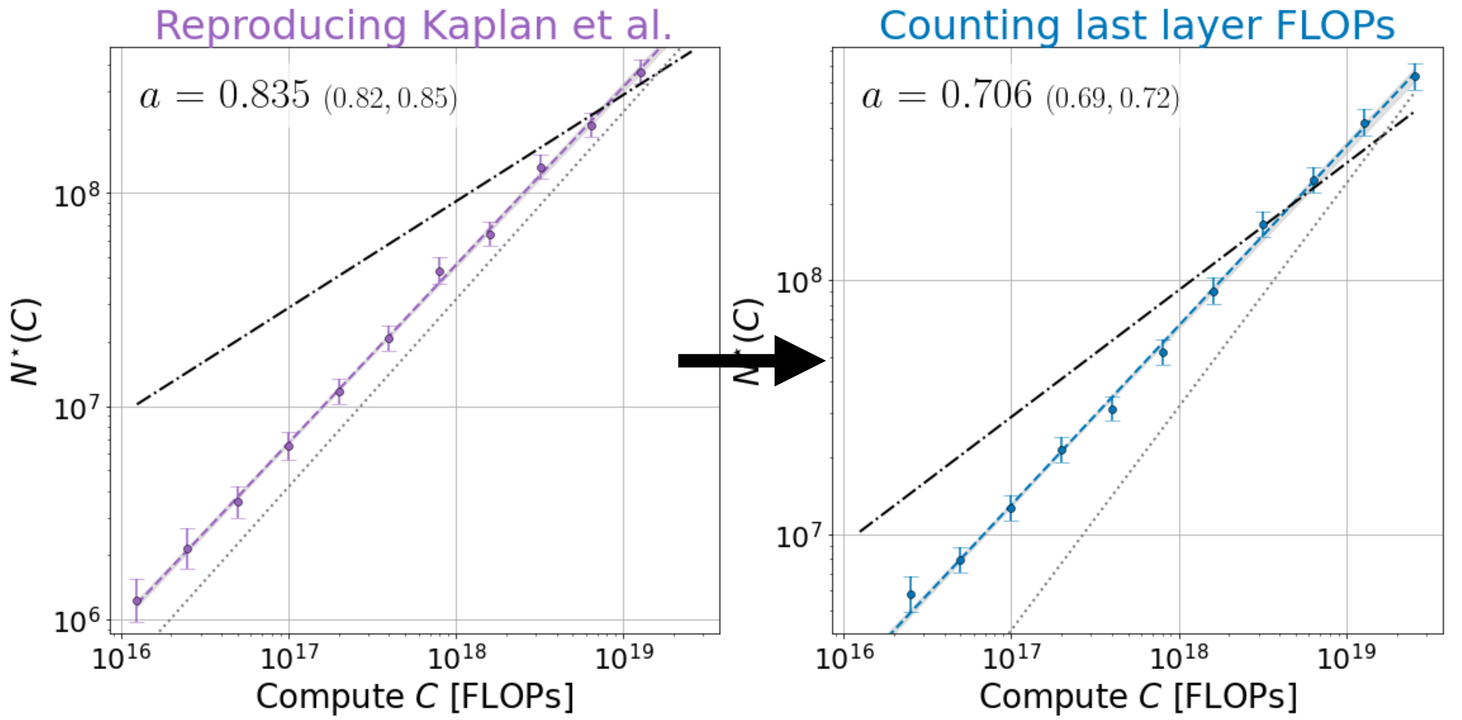
Reproducing Kaplan et al.



Three reasons

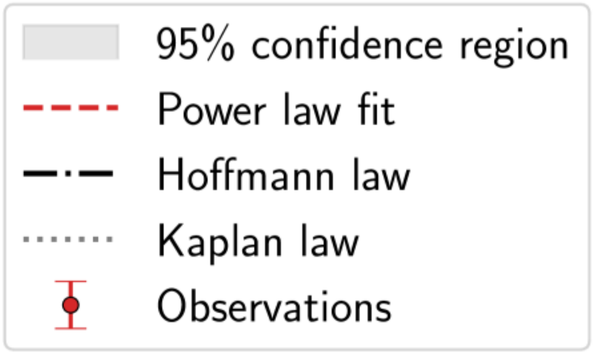
for the discrepancy



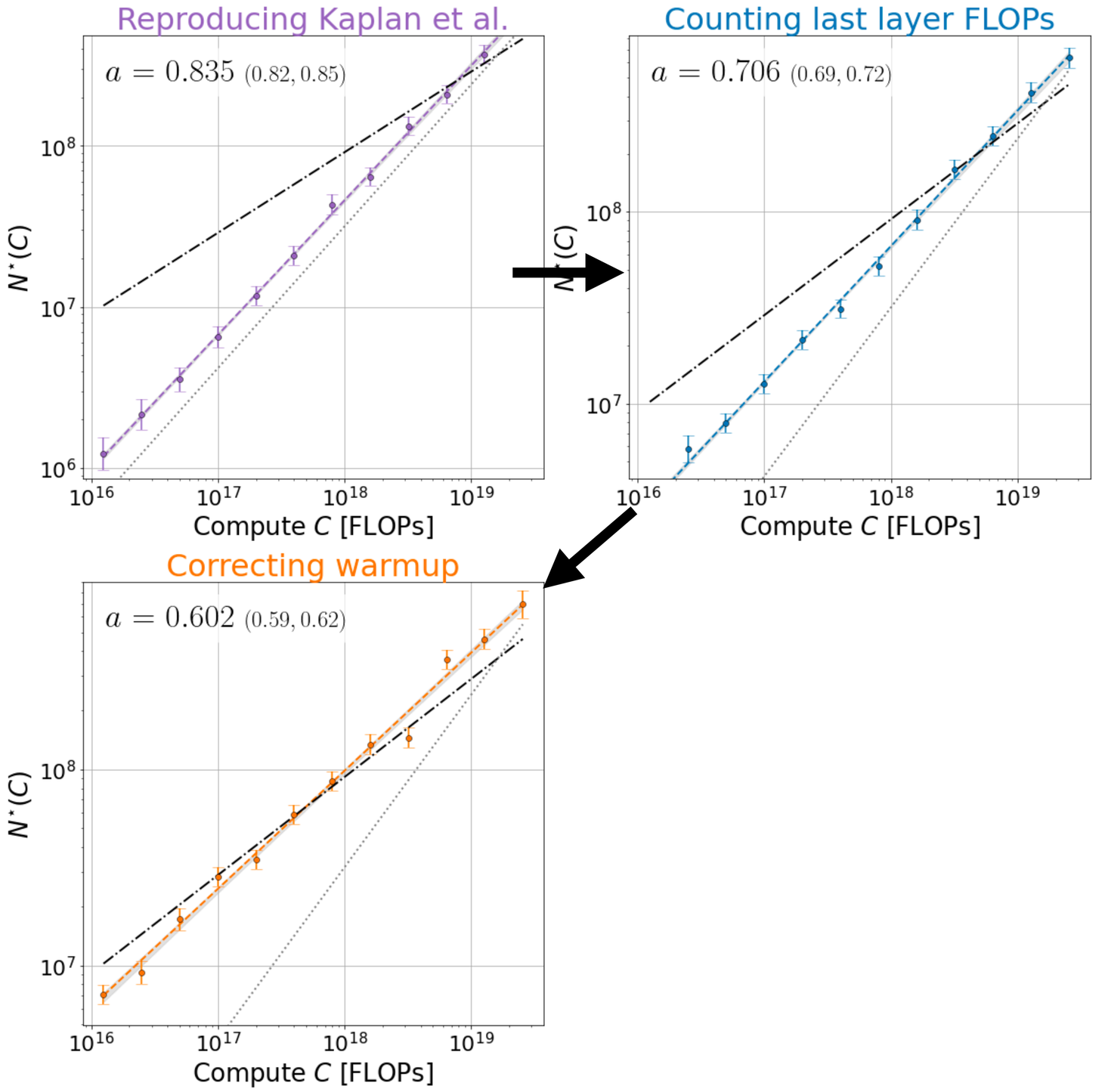


1. Definition of model size

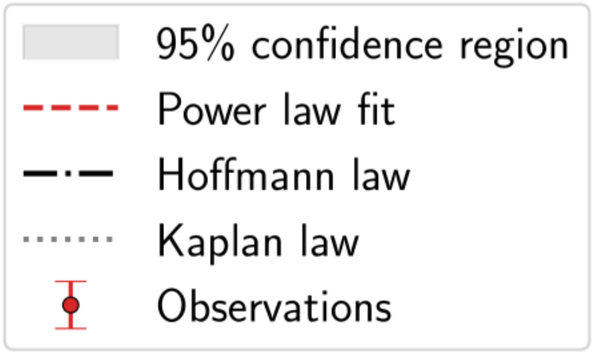
Three reasons
for the discrepancy



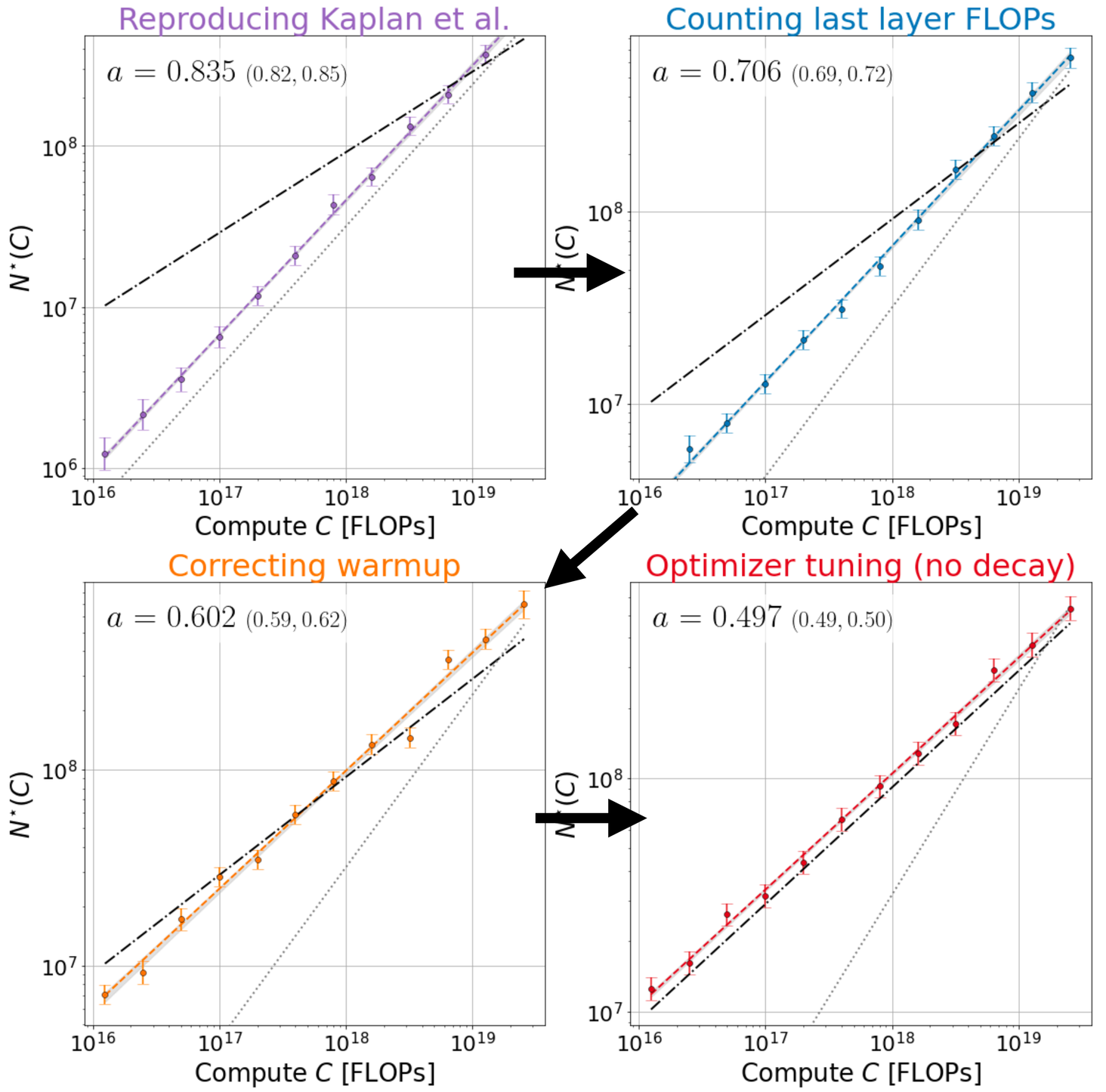
1. Definition of model size
2. Warmup duration



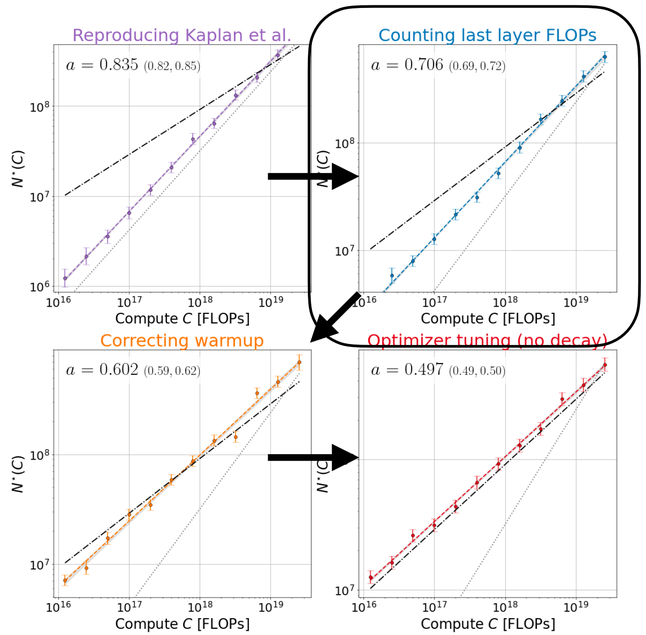
Three reasons
for the discrepancy



1. Definition of model size
2. Warmup duration
3. Optimizer tuning

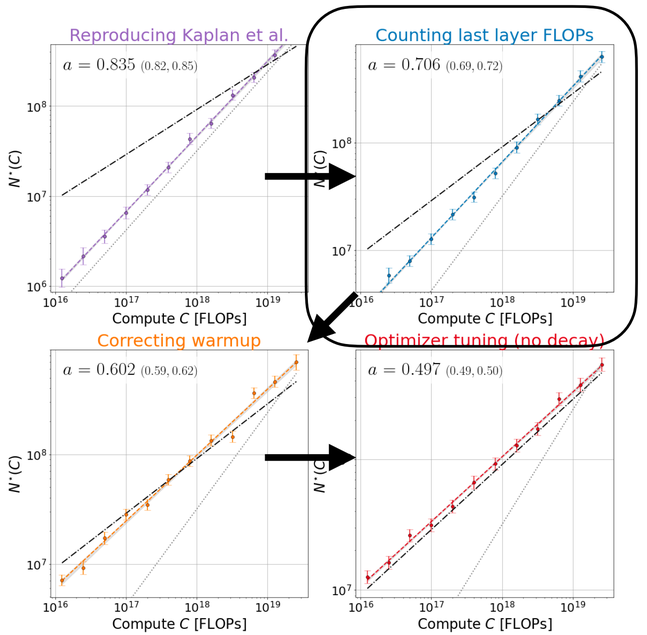


Counting last layer FLOPs



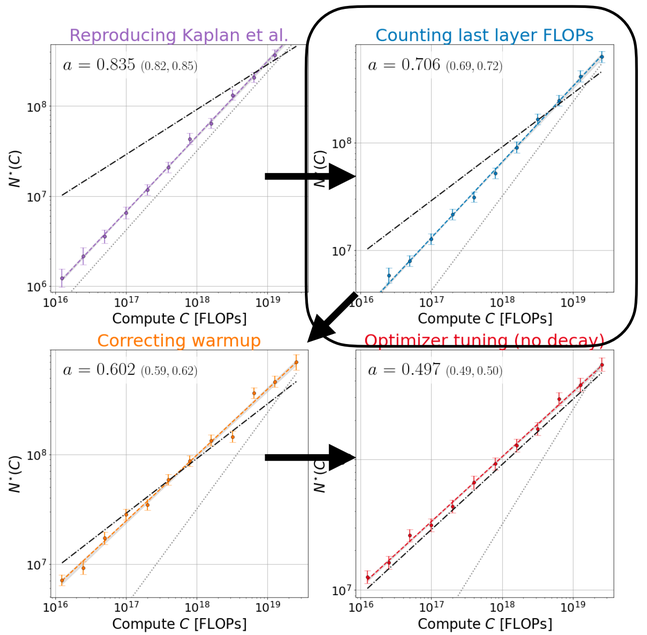
Counting last layer FLOPs

All parameters in linear layers	$\left(3d_{\mathrm{FF}} + 4d\right) dl + dv$
All parameters in linear layer *except the last one*	$\left(3d_{\mathrm{FF}} + 4d\right) dl$

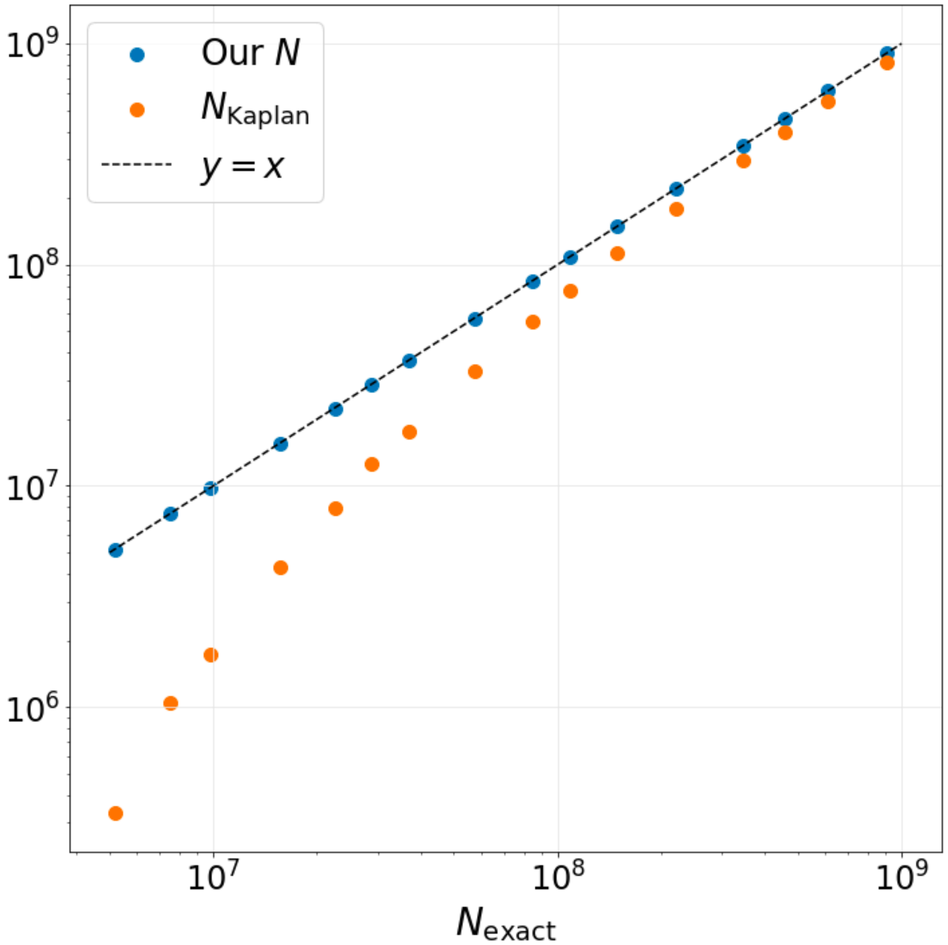


Counting last layer FLOPs

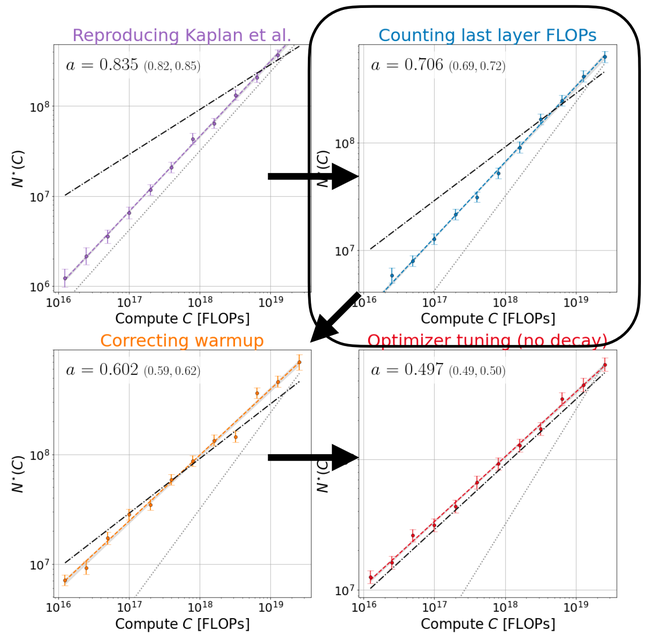
All parameters in linear layers	$\left(3d_{\text{FF}} + 4d\right) dl + dv$
All parameters in linear layer *except the last one*	$\left(3d_{\text{FF}} + 4d\right) dl$



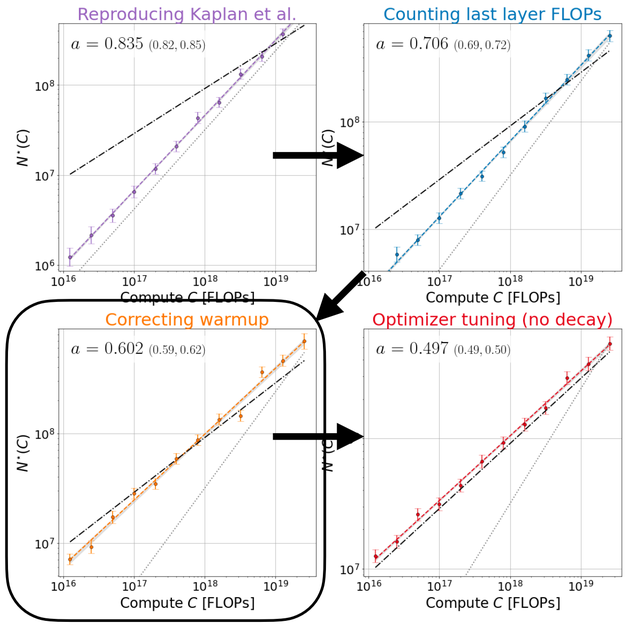
Counting last layer FLOPs



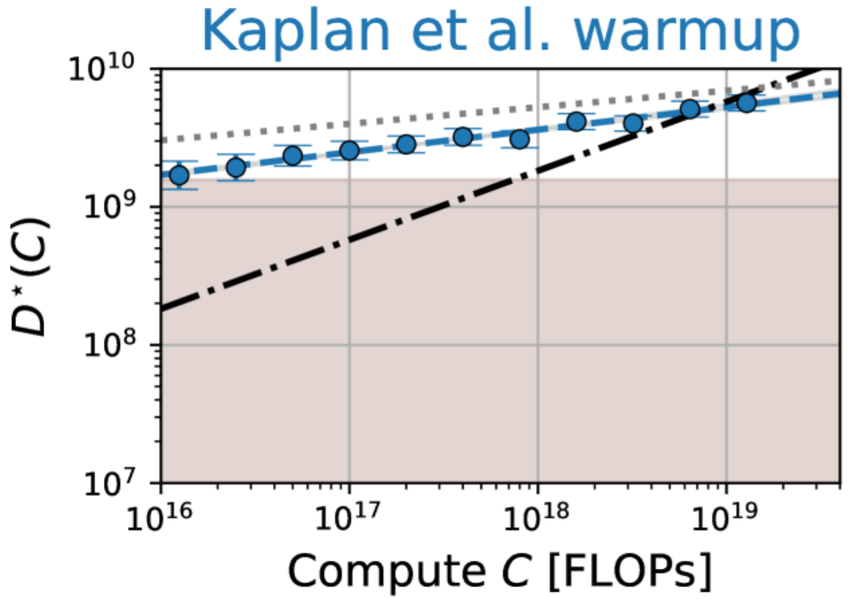
All parameters in linear layers	$\left(3d_{\text{FF}} + 4d\right) dl + dv$
All parameters in linear layer *except the last one*	$\left(3d_{\text{FF}} + 4d\right) dl$

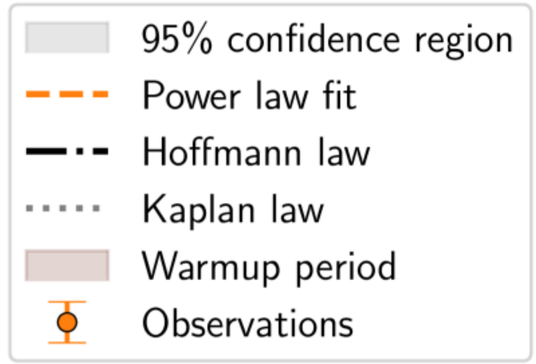


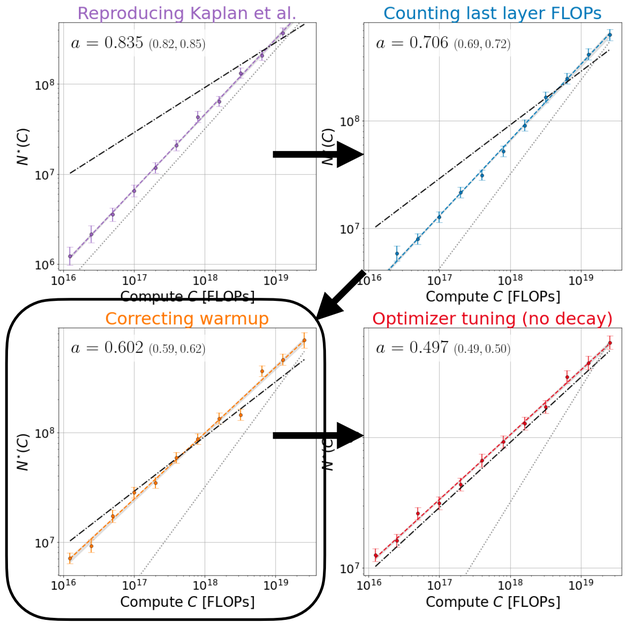
Warmup duration



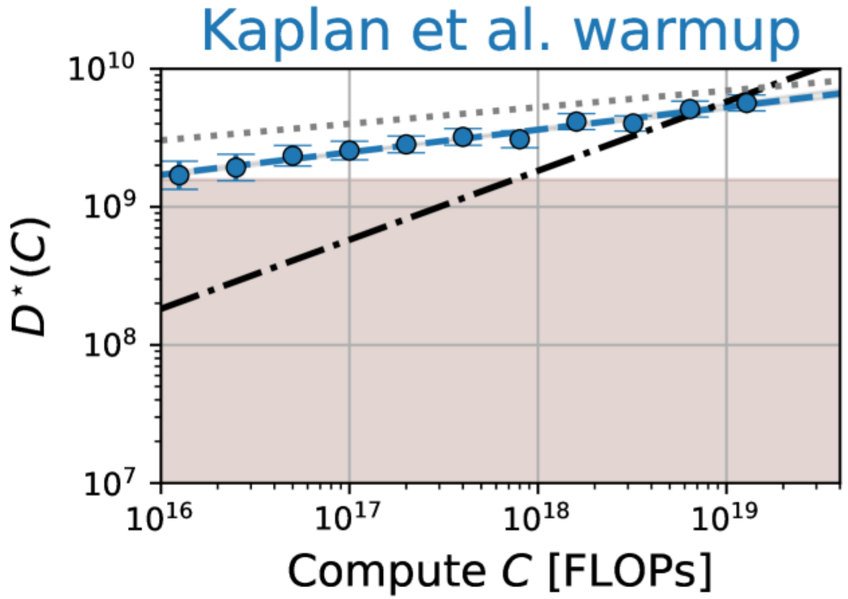
Warmup duration

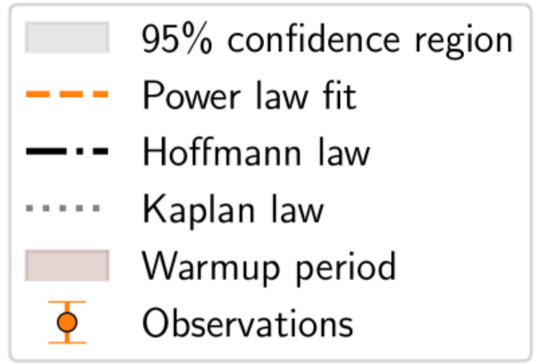


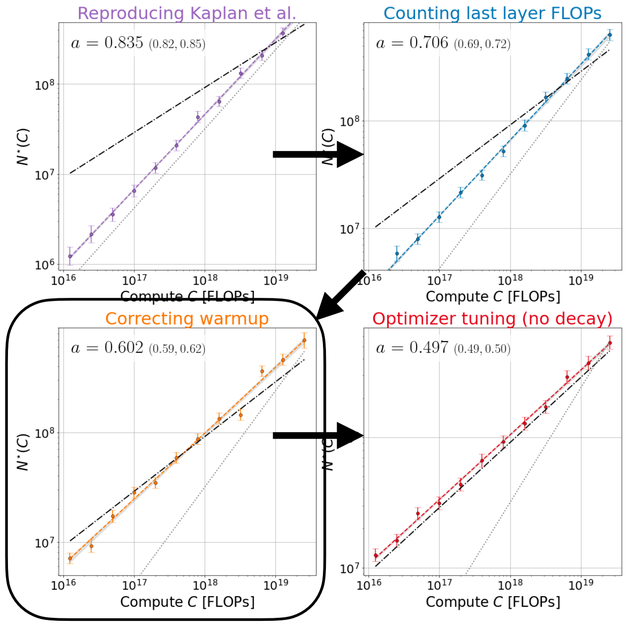




Warmup duration

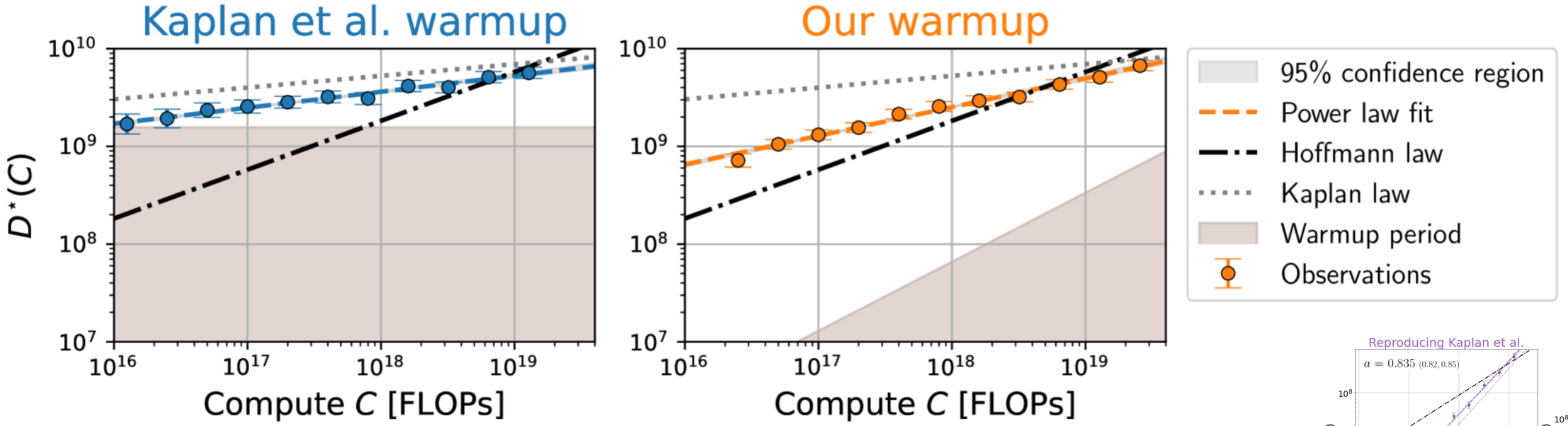


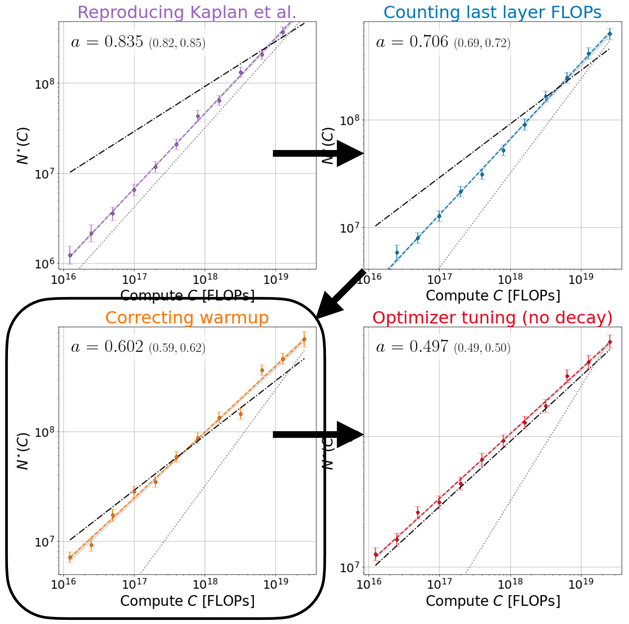




Simple heuristic: warmup tokens = N

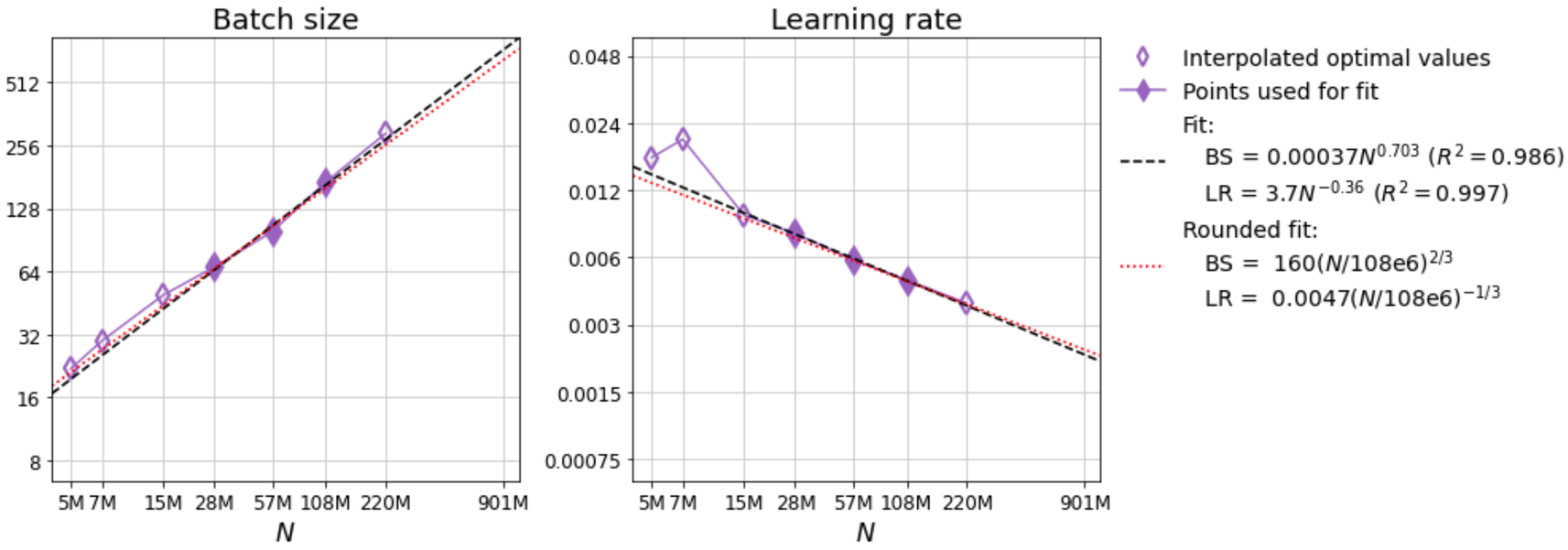
Warmup duration

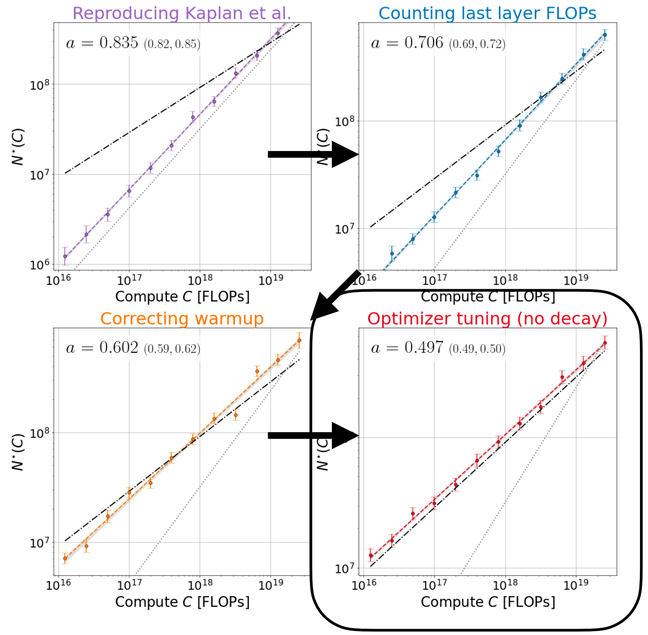




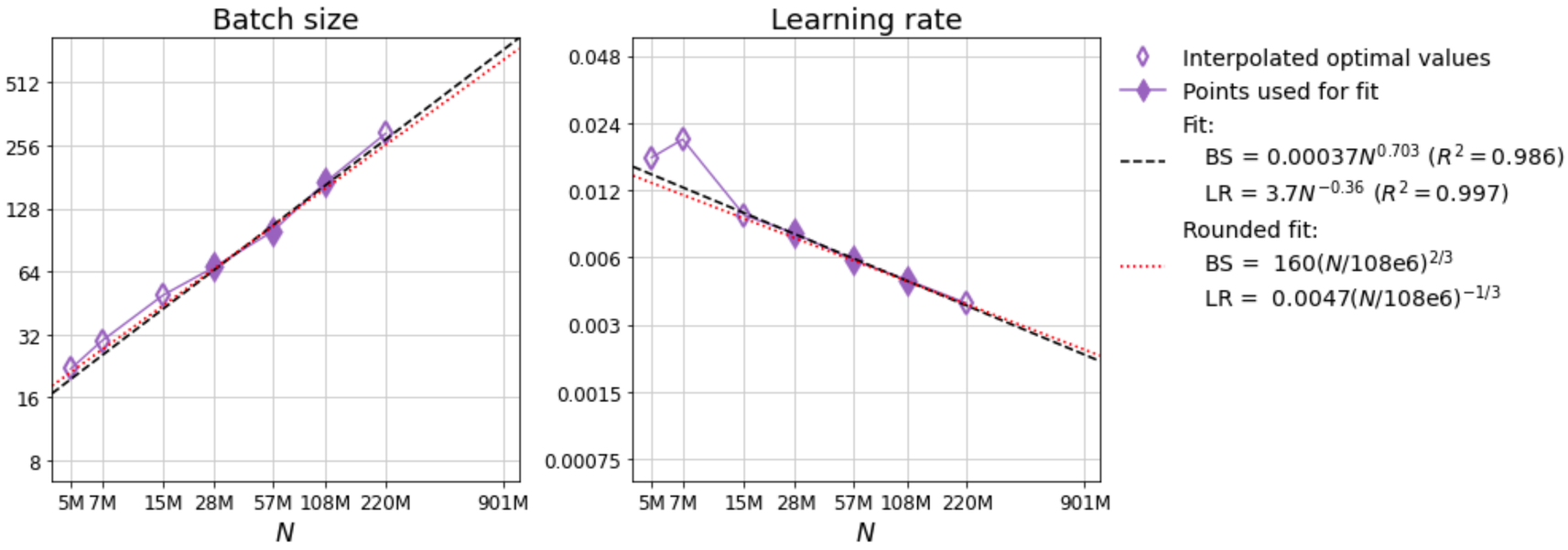
Simple heuristic: warmup tokens = N

Optimizer tuning

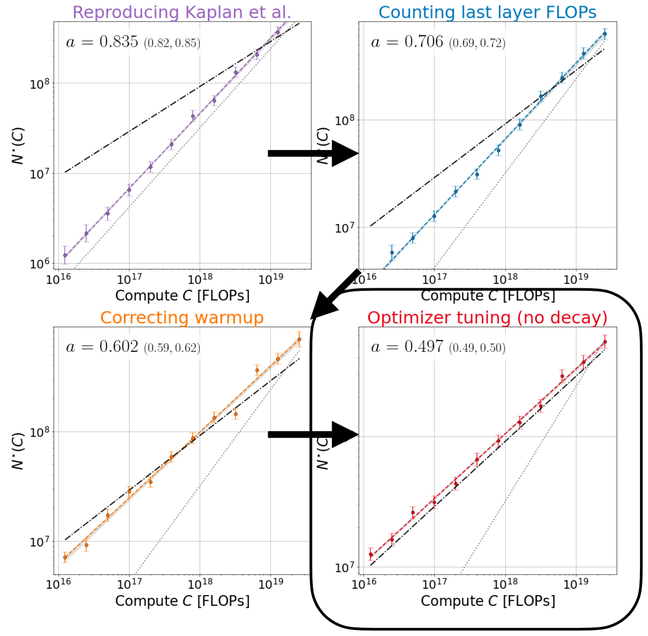




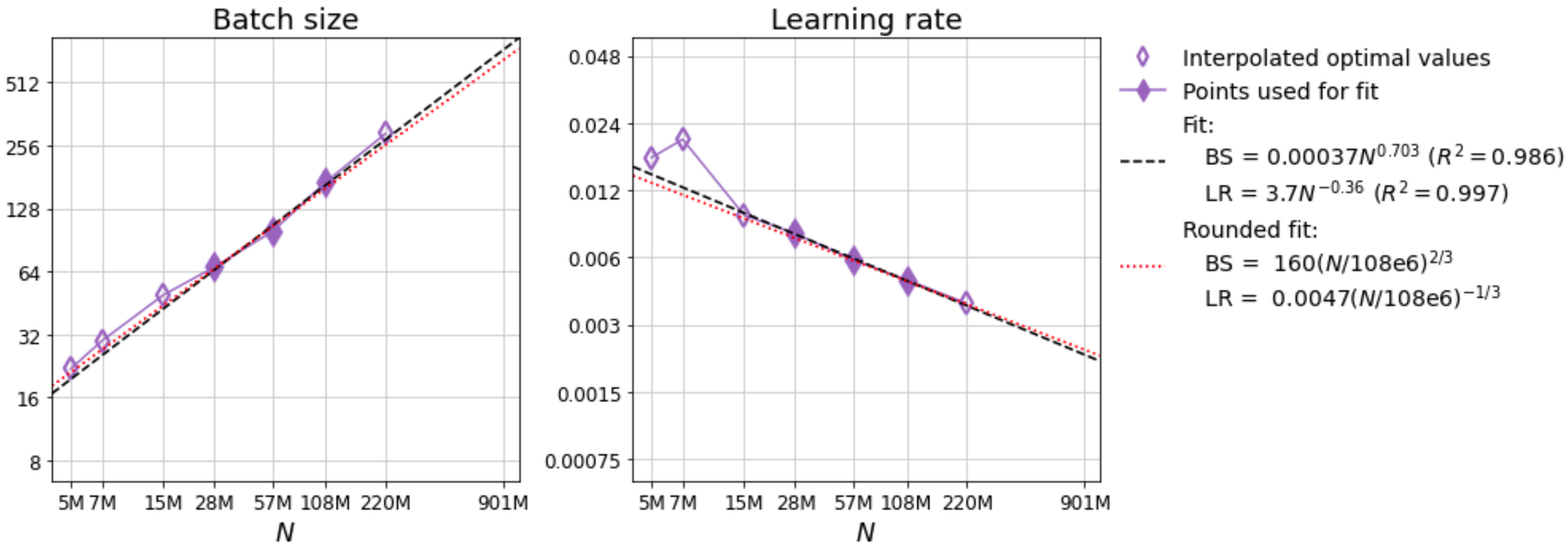
Optimizer tuning



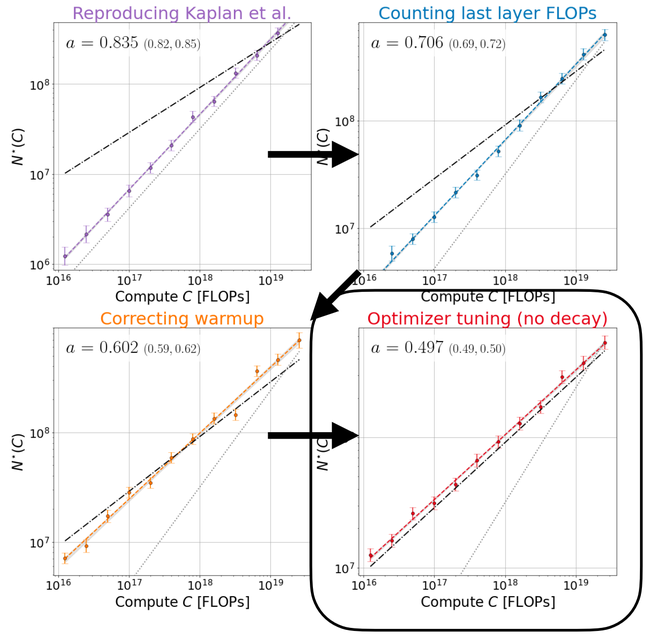
- We sweep LR and BS on 7 models (5M to 220M) with 7 values for LR and 7 values for BS



Optimizer tuning

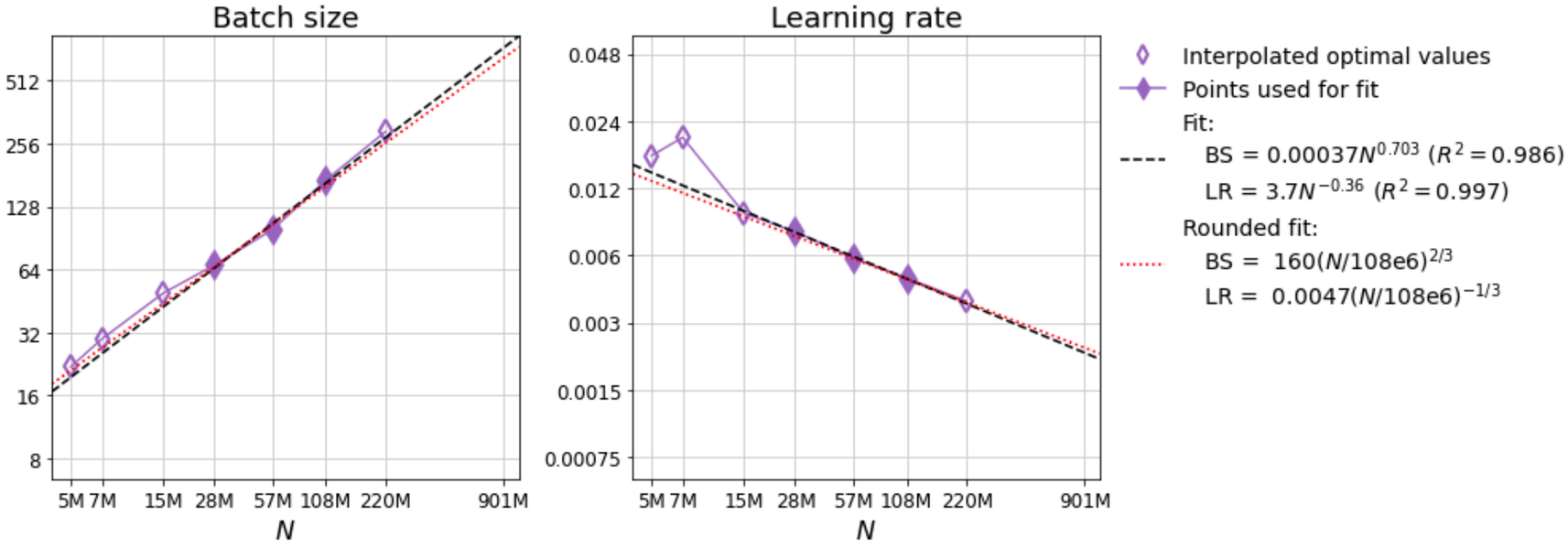


- We sweep LR and BS on 7 models (5M to 220M) with 7 values for LR and 7 values for BS
- Inspired by DeepSeek, we fit a scaling law for BS and LR

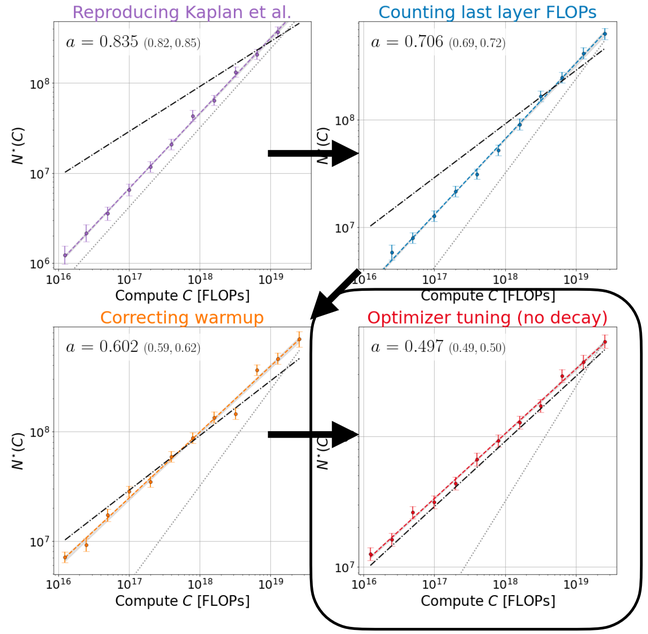


DeepSeek.
Deepseek LLM: Scaling open-source language models with longtermism.
arXiv:2401.02954, 2024.

Optimizer tuning



- We sweep LR and BS on 7 models (5M to 220M) with 7 values for LR and 7 values for BS
- Inspired by DeepSeek, we fit a scaling law for BS and LR
- It is crucial to choose higher values β_2 in AdamW for small BS

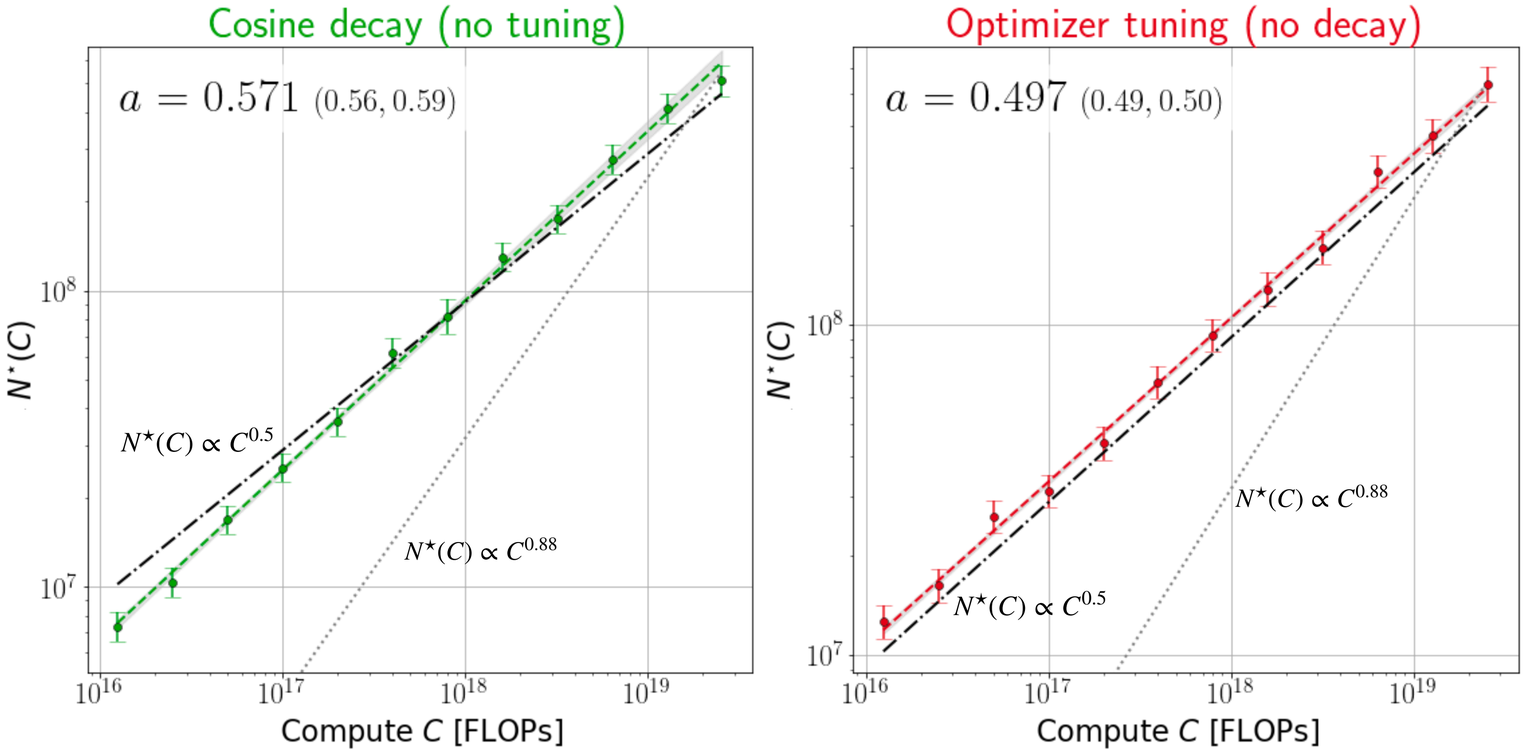


DeepSeek.
Deepseek LLM: Scaling open-source language models with longtermism.
arXiv:2401.02954, 2024.

Additional analysis

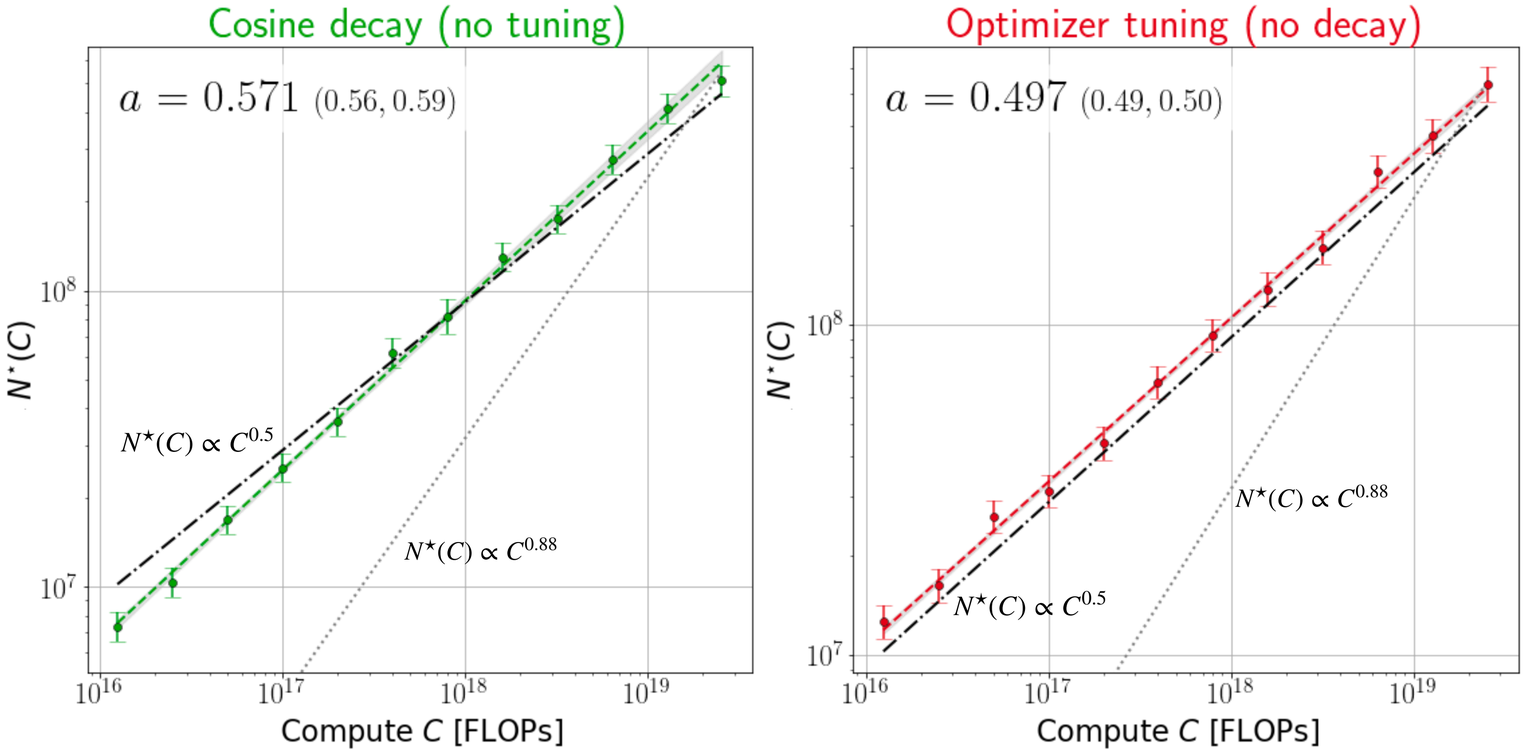
Additional analysis

Contrary to Hoffmann et al.'s hypothesis, LR decay is not crucial for correct scaling



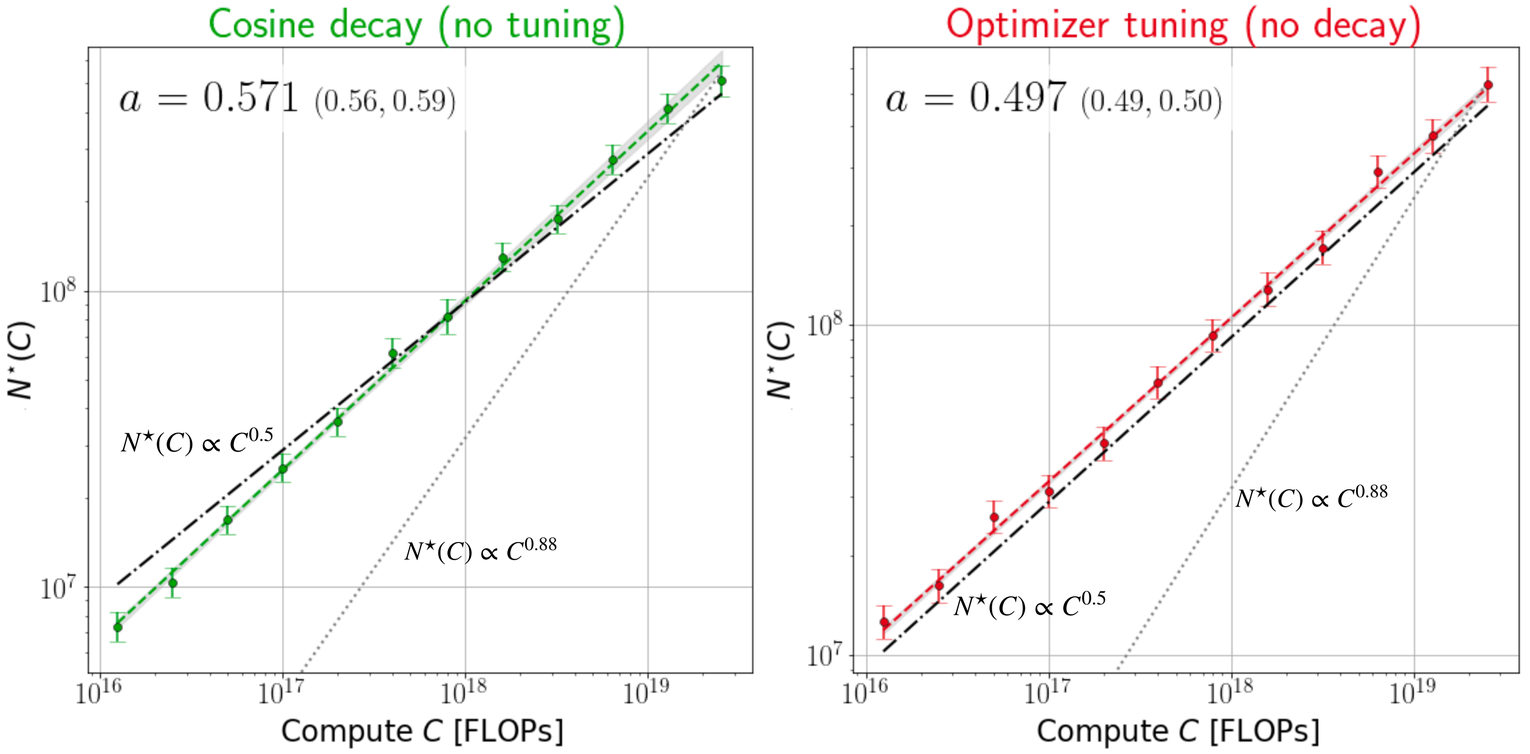
Additional analysis

Contrary to Hoffmann et al.'s hypothesis, LR decay is not crucial for correct scaling



Additional analysis

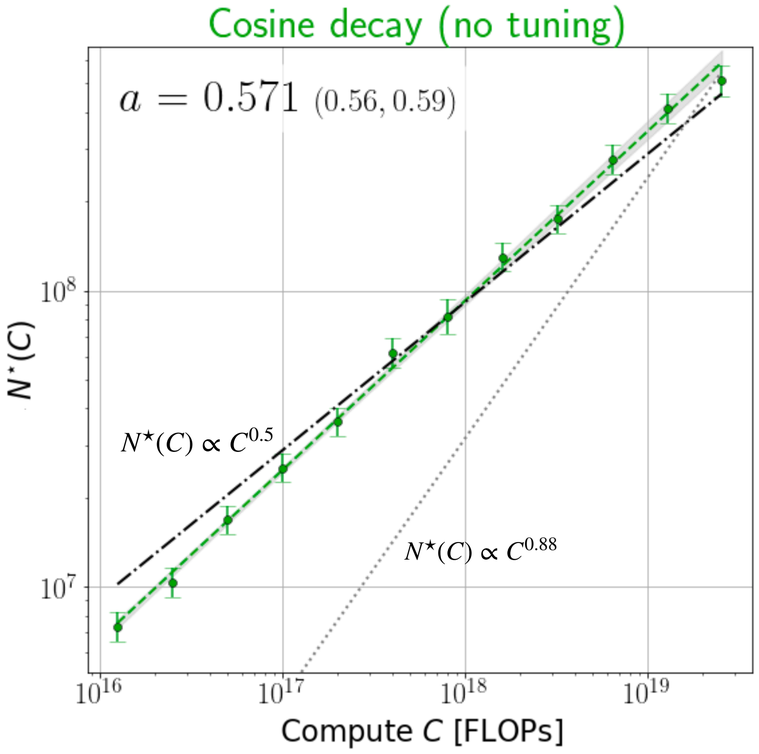
Contrary to Hoffmann et al.'s hypothesis, LR decay is not crucial for correct scaling

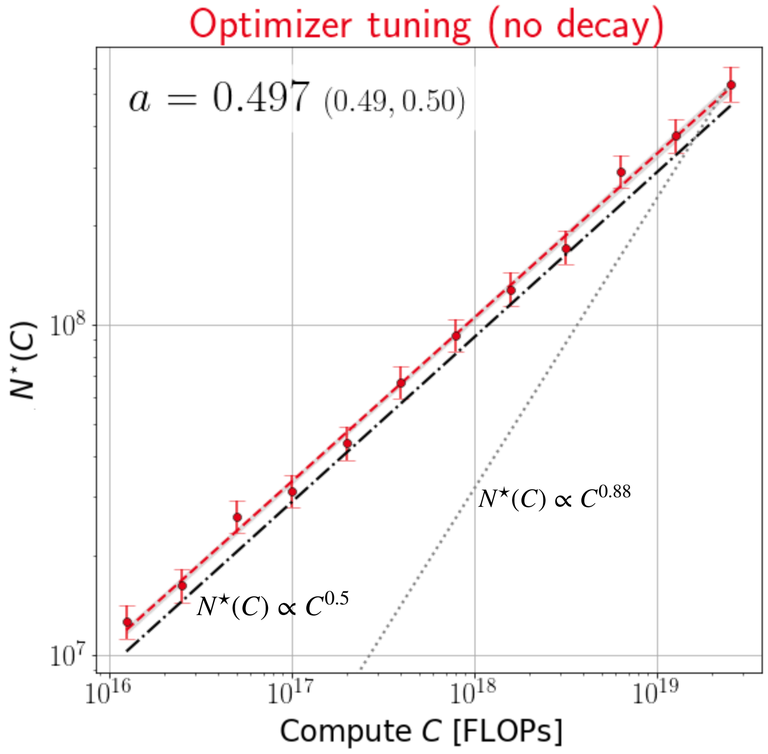


We can reproduce the post-processed exponent $a \approx 0.73$ by tuning the optimizer but not correcting the FLOP counts and warmup

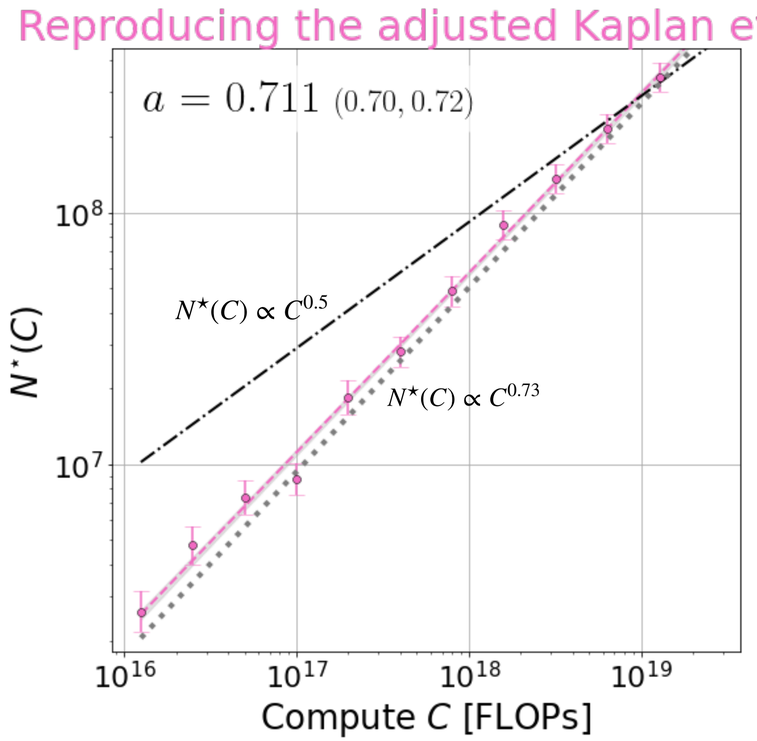
Additional analysis

Contrary to Hoffmann et al.'s hypothesis, LR decay is not crucial for correct scaling





We can reproduce the post-processed exponent $a \approx 0.73$ by tuning the optimizer but not correcting the FLOP counts and warmup



Thank you!

Code, data and checkpoints available online

Tomer Porian

Mitchell Wortsman

Jenia Jitsev

Ludwig Schmidt

Yair Carmon

TEL AVIV UNIVERSITY

UNIVERSITY of WASHINGTON

JÜLICH FORSCHUNGSZENTRUM