Switch Transformers

Scaling to Trillion Parameter Models With Simple and Efficient Sparsity

arXiv:210103961

Sam Foreman

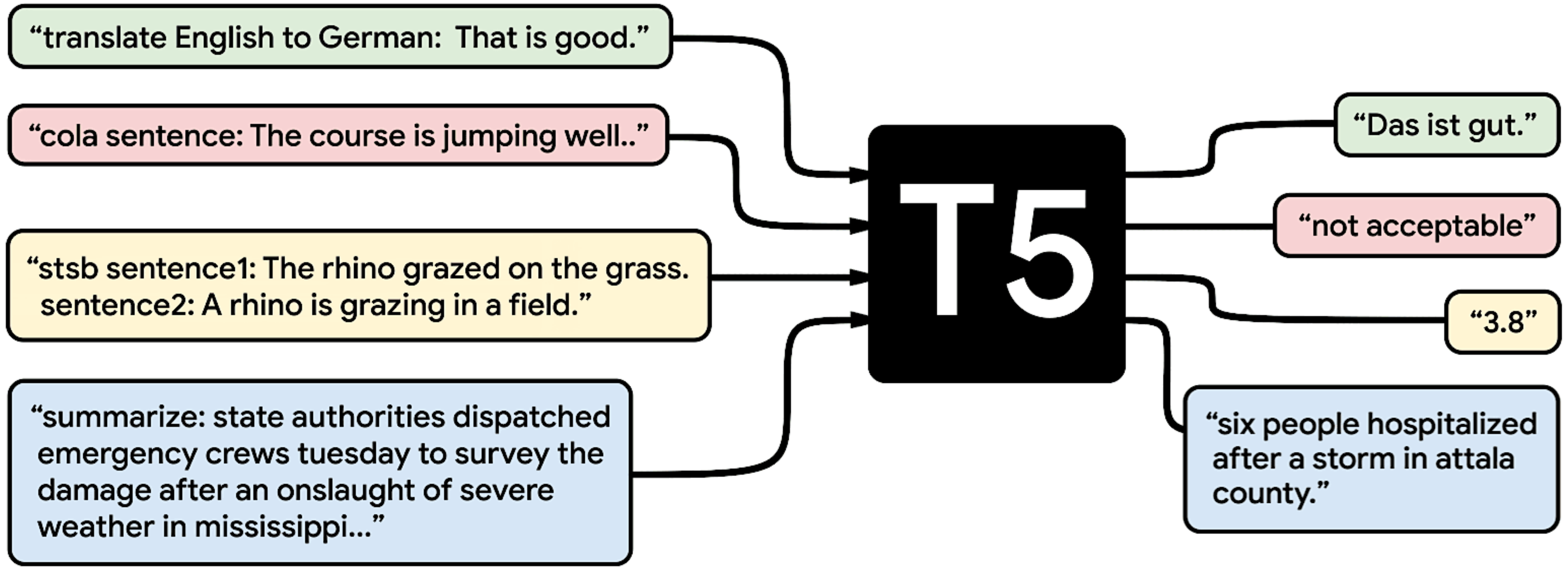
June, 2021

Argonne
NATIONAL LABORATORY

Switch Transformers

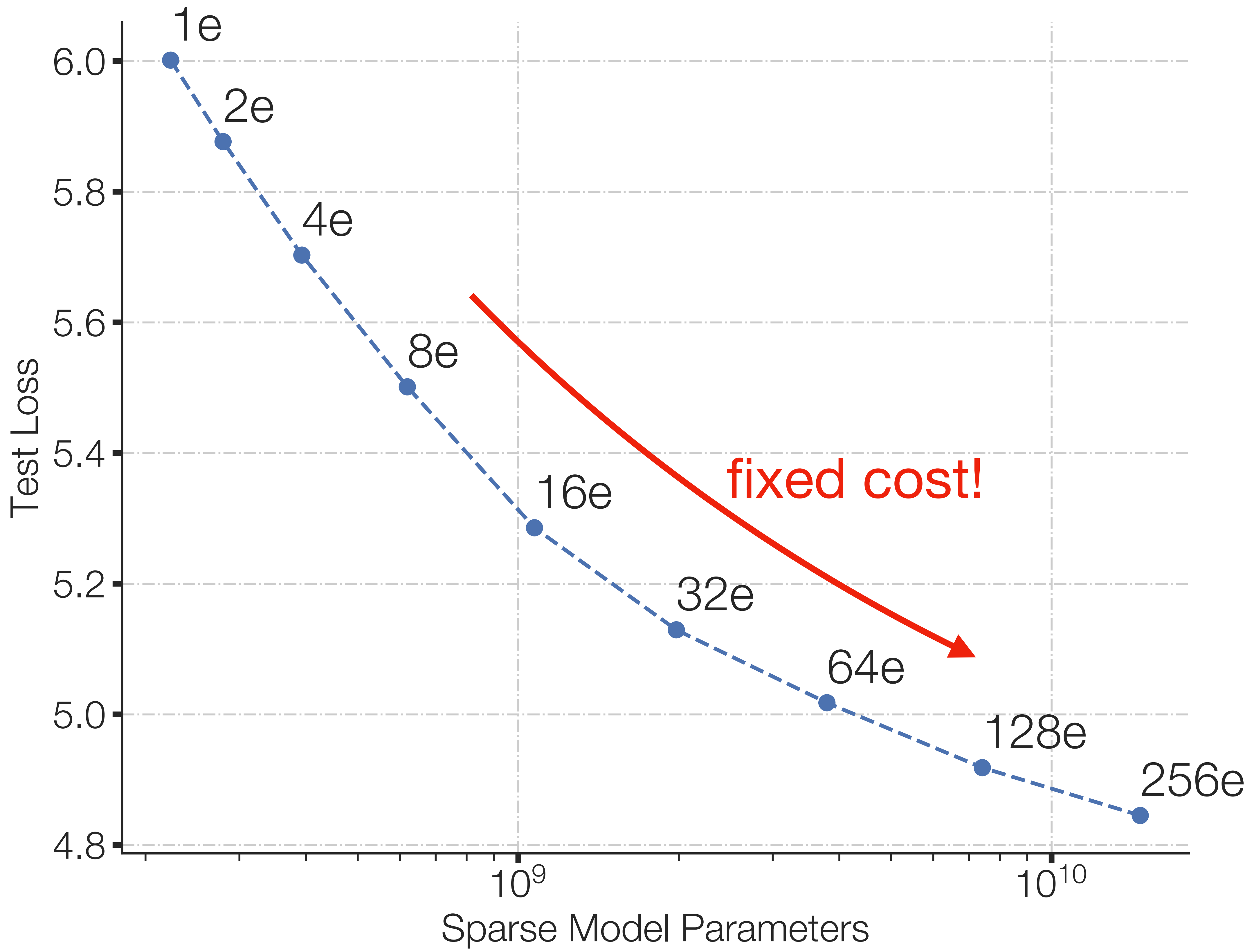
More parameters, same cost!

- **Big Idea:** We don't *always* need to know *everything*.
- **Mixture-of-Experts** (MOE): provides an efficient mechanism for scaling up the number of parameters while keeping the total training cost fixed.
  - Clever engineering avoids training instability
- **1.6 trillion parameters** (most to date), **7x speedup** over T5-BASE [1.]

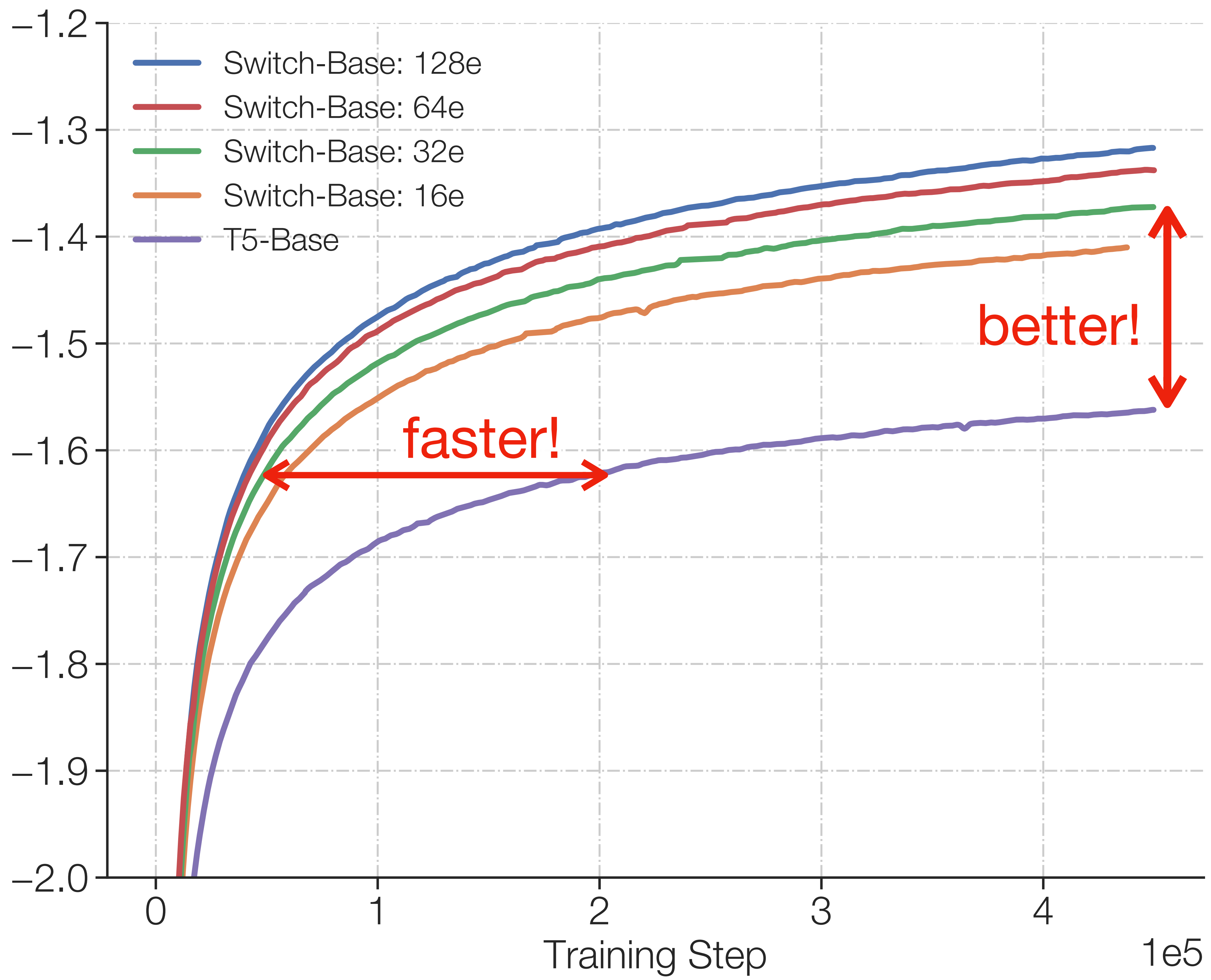

"translate English to German: That is good."
"cola sentence: The course is jumping well.."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

[1.] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer arXiv:1910.10683

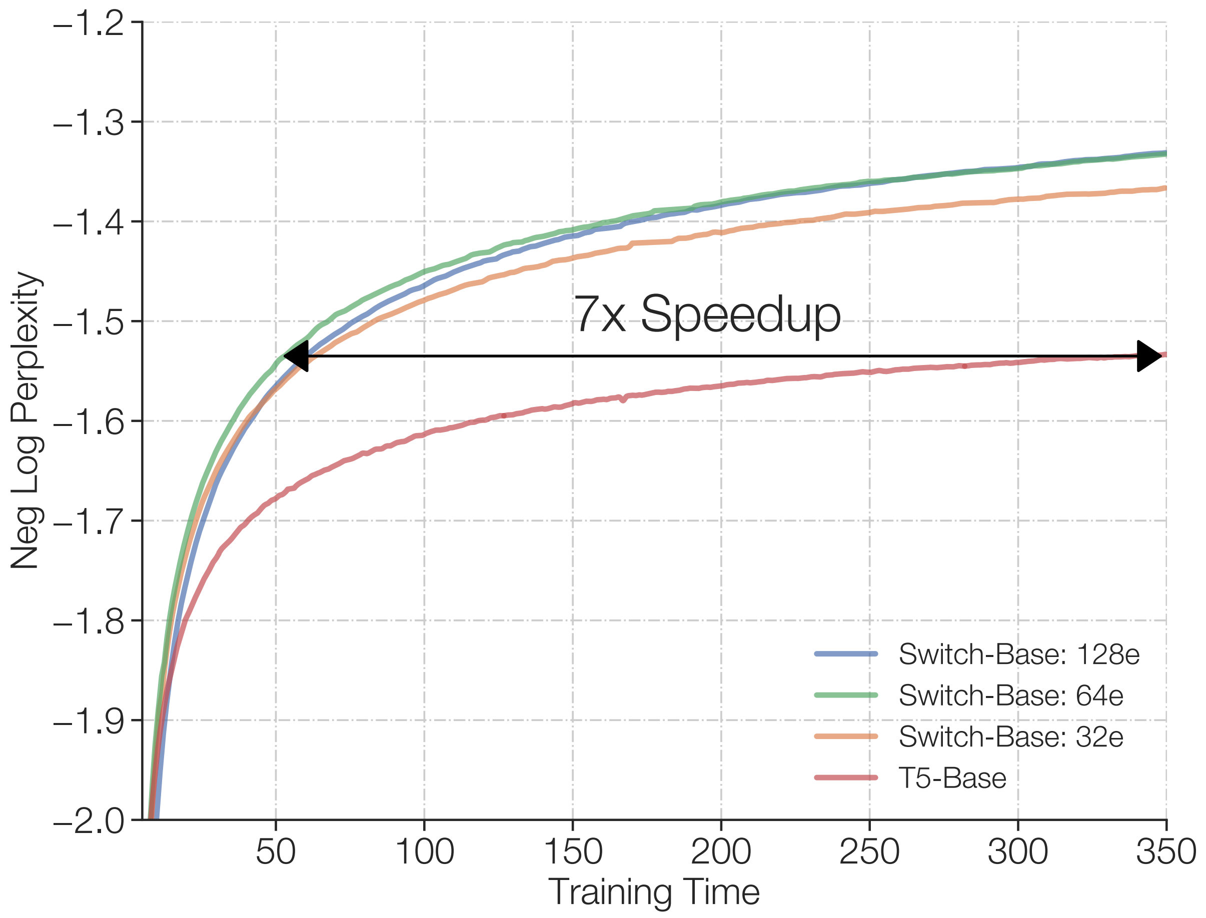
[2.] https://www.youtube.com/watch?v=iAR8LkkMMIM



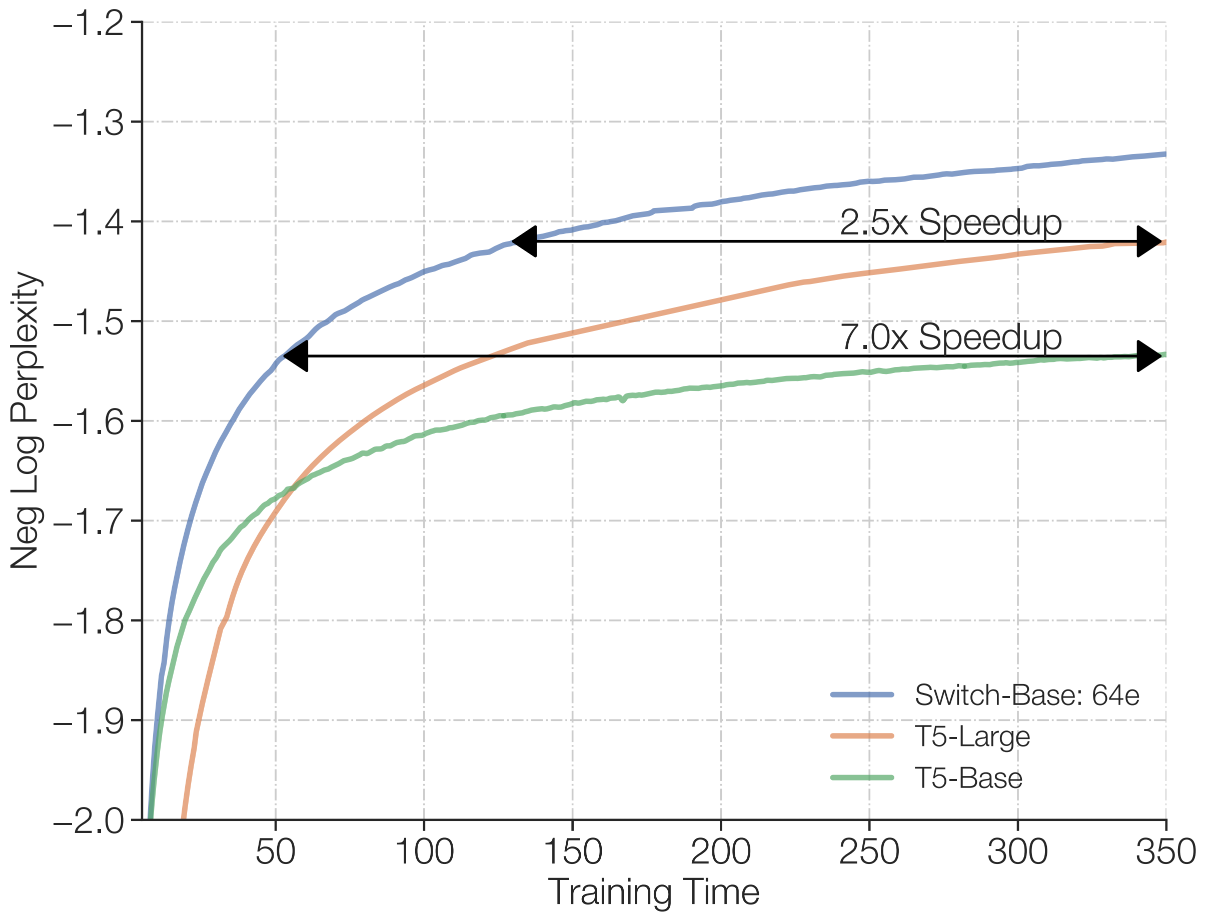
1e
2e
4e
8e
16e
32e
64e
128e
256e
fixed cost!
Test Loss
6.0
5.8
5.6
5.4
5.2
5.0
4.8
10^9
10^{10}
Sparse Model Parameters



Switch-Base: 128e
Switch-Base: 64e
Switch-Base: 32e
Switch-Base: 16e
T5-Base
better!
faster!
−1.2
−1.3
−1.4
−1.5
−1.6
−1.7
−1.8
−1.9
−2.0
0
1
2
3
4
Training Step
1e5



7x Speedup
Neg Log Perplexity
−1.2
−1.3
−1.4
−1.5
−1.6
−1.7
−1.8
−1.9
−2.0
50
100
150
200
250
300
350
Training Time
Switch-Base: 128e
Switch-Base: 64e
Switch-Base: 32e
T5-Base

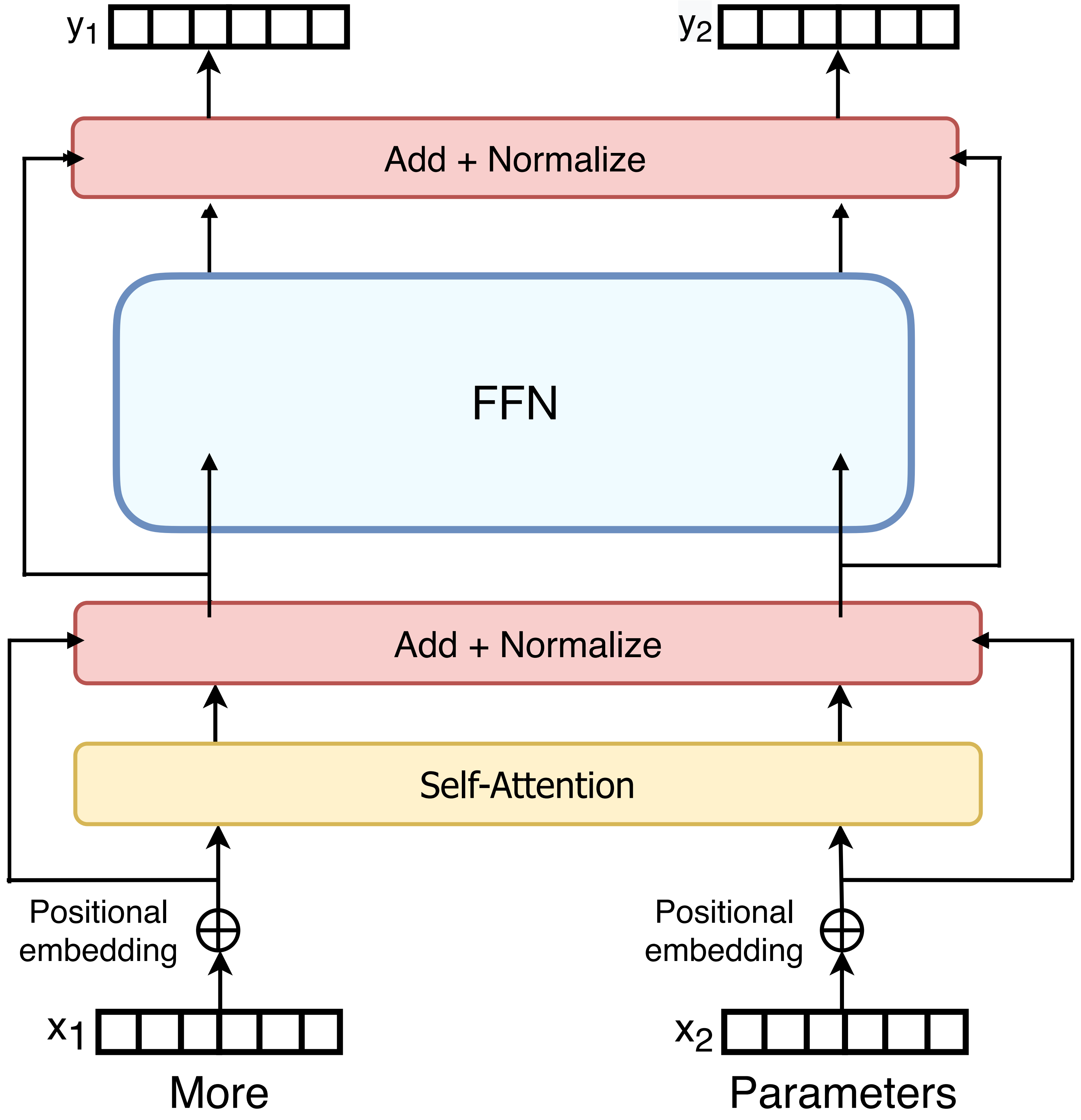


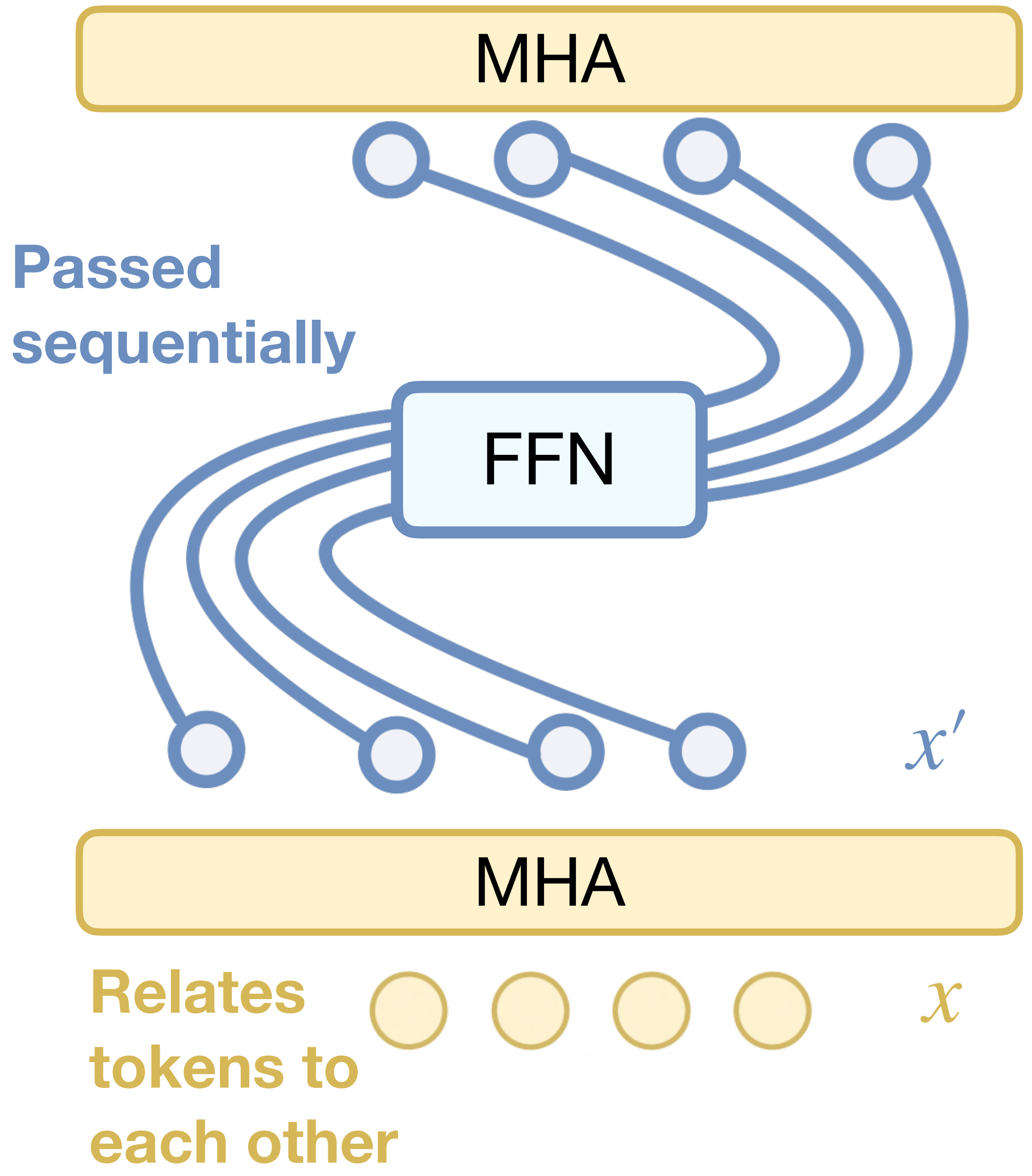
2.5x Speedup
7.0x Speedup
Neg Log Perplexity
−1.2
−1.3
−1.4
−1.5
−1.6
−1.7
−1.8
−1.9
−2.0
50
100
150
200
250
300
350
Training Time
Switch-Base: 64e
T5-Large
T5-Base

Transformer Architecture

- **Multi-Head Attention (MHA):**
  - Aggregates information from sequences
  - Relates tokens to each other

- **Feed Forward (FFN):**
  - Aggregates outputs from multiple heads
  - Tokens in sequence are passed sequentially
  - For a given token and its representation in this layer, what is the best representation in the next layer?





Transformer Architecture

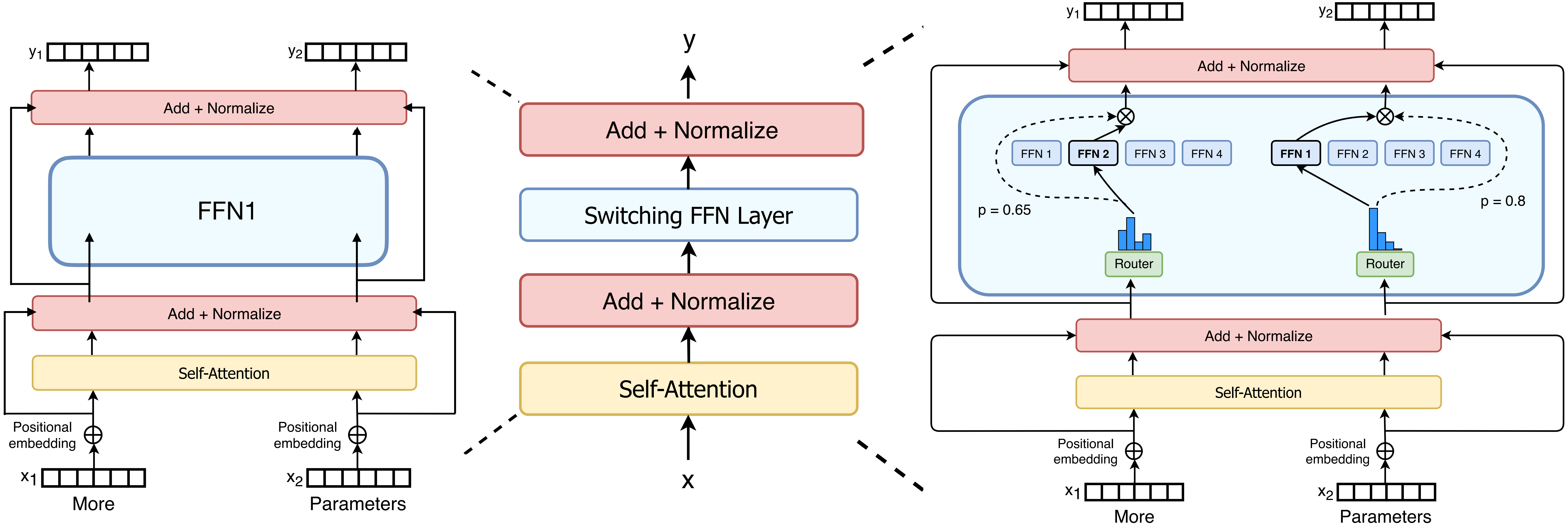
Using multiple FFNs

- **Multi-Head Attention (MHA):**
  - Aggregates information from sequences
  - Relates tokens to each other

- **Feed Forward (FFN):**
  - Tokens in sequence are passed sequentially
  - For a given token and its representation in this layer, what is the best representation in the next layer?

Passed sequentially

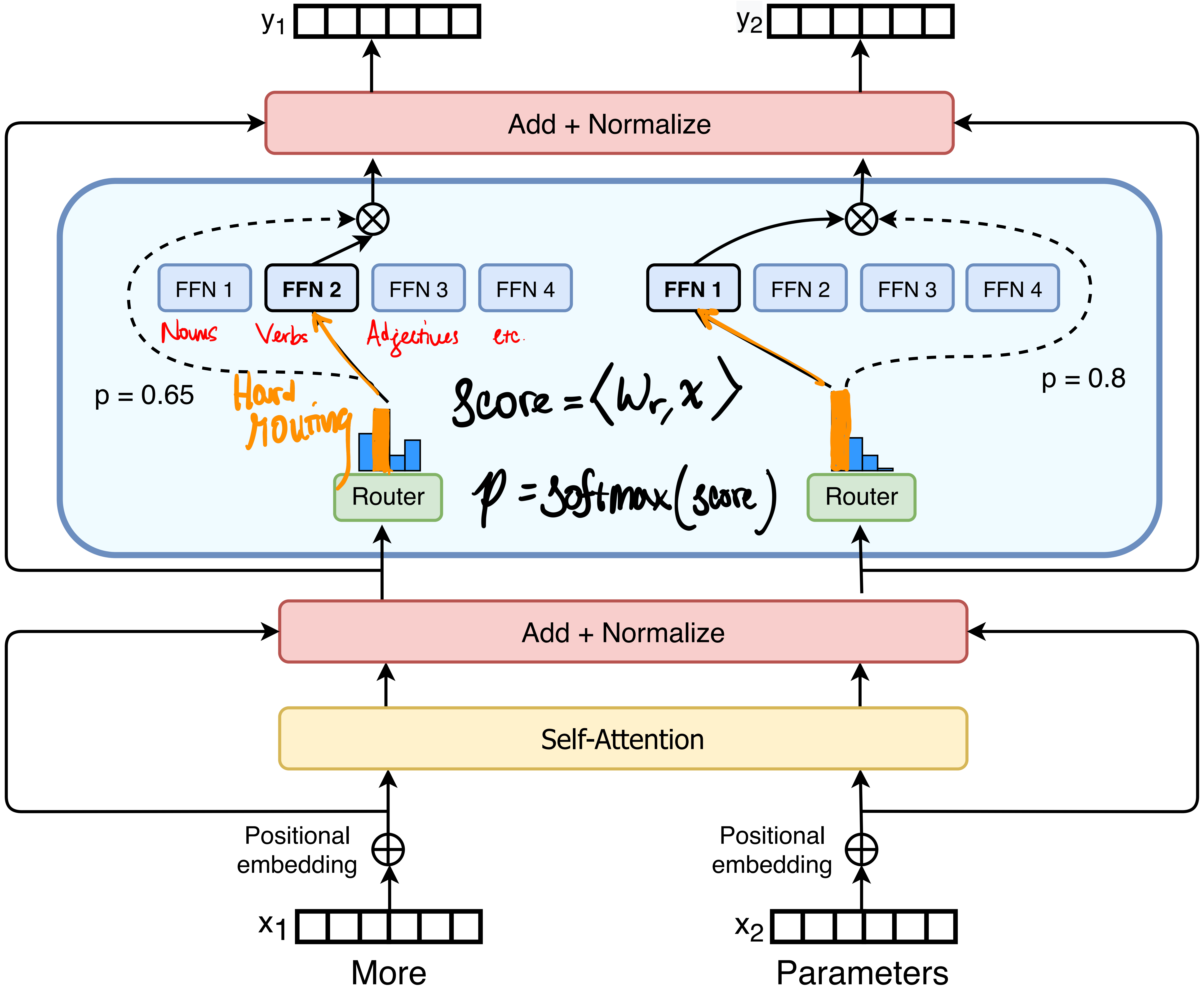
Relates tokens to each other



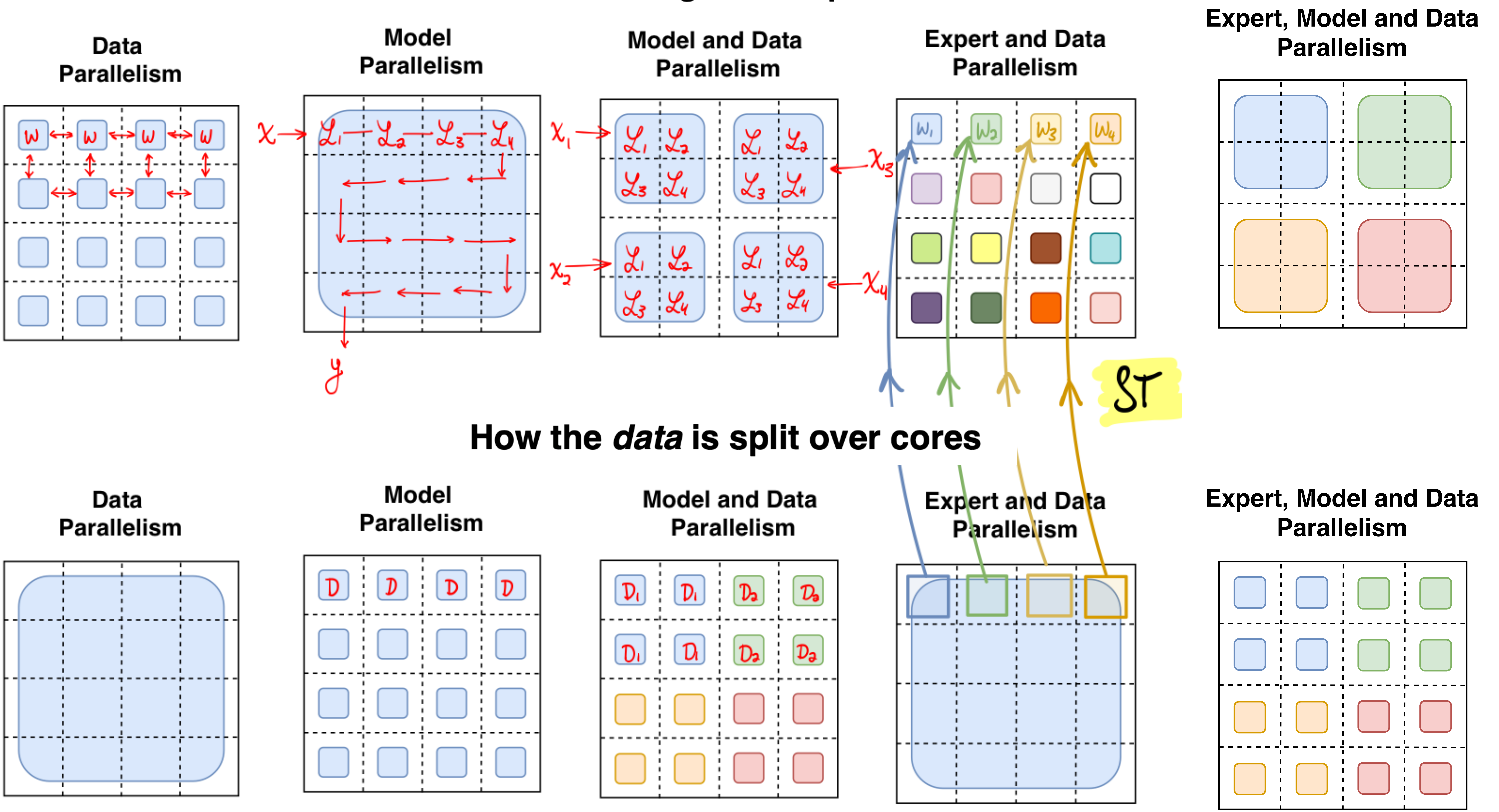
Switch Layer

Learning to Route inputs

- We don't need *all* the model's information for a particular input
  - **Use different parameters for different inputs**
- Replace *single* FFN with *multiple* FFNs, referred to as *experts.*
  - Each token is passed through exactly one FFN
  - Requires more memory
  - Constant FLOPs
- **Switch Layer:** Learns how to route each token to the most suitable expert.
- Hard routing introduces sparsity!



How the *model weights* are split over cores



Data Parallelism

Model Parallelism

Model and Data Parallelism

Expert and Data Parallelism

Expert, Model and Data Parallelism

How the *data* is split over cores

Data Parallelism

Model Parallelism

Model and Data Parallelism

Expert and Data Parallelism

Expert, Model and Data Parallelism

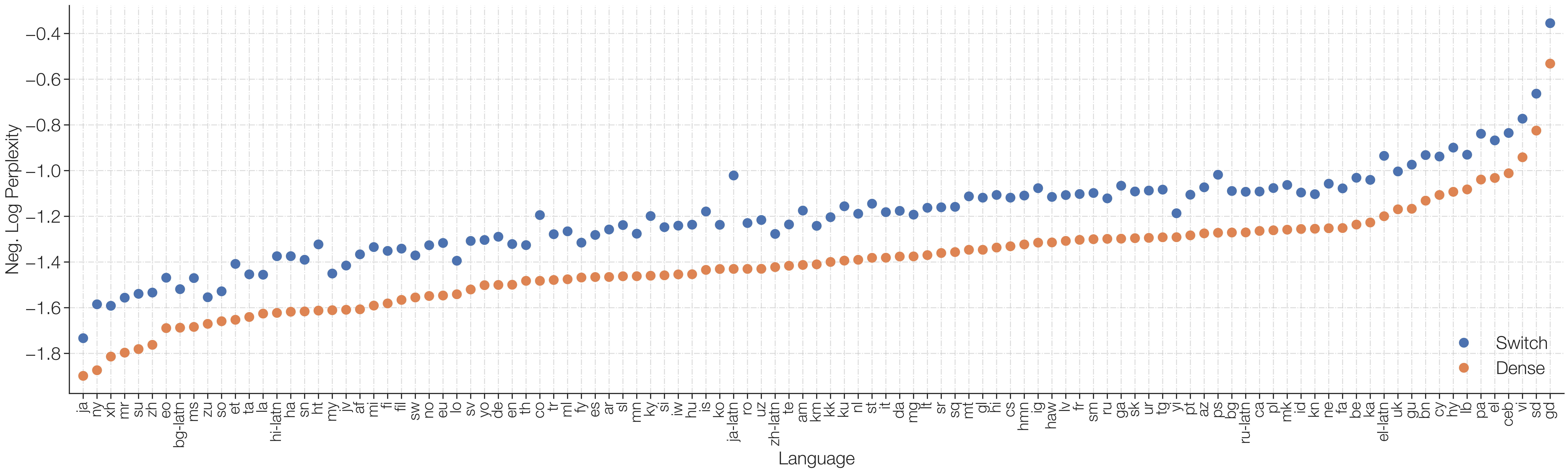
Stabilizing Training

- **Selective precision with large sparse models.**
  - Cast input to “switch” router up to float32 precision.
  - Cast output back to bfloat16
- **Smaller parameter initialization for stability.**
  - Initialize weight matrices from truncated normal with $\mu = 0,\ \sigma = \sqrt{s/n}$, and s is a hyperparameter
    - Recommend scaling default value ($s = 1.0$) by $\sim 1/10.$
- **Regularizing large sparse models**
  - Significantly increase dropout probability inside the experts

Downstream Results

Multilingual Learning

- Outperforms T5 Base model *on all 101 languages!*


Neg. Log Perplexity
Switch
Dense
ja
ny
xh
mr
su
zh
eo
bg-latn
ms
zu
so
et
ta
la
hi-latn
ha
sn
ht
my
jv
af
mi
fi
fil
sw
no
eu
lo
sv
yo
de
en
th
co
tr
ml
fy
es
ar
sl
mn
ky
si
iw
hu
is
ko
ja-latn
ro
uz
zh-latn
te
am
km
kk
ku
nl
st
it
da
mg
lt
sr
sq
mt
gl
hi
cs
hmn
ig
haw
lv
fr
sm
ru
ga
sk
ur
tg
yi
pt
az
ps
bg
ru-latn
ca
pl
mk
id
kn
ne
fa
be
ka
el-latn
uk
gu
bn
cy
hy
lb
pa
el
ceb
vi
sd
gd
Language