FlexOlmo:
Open Language Models for Flexible Data Use

Sewon Min

sewonmin.com

Berkeley UNIVERSITY OF CALIFORNIA BAIR Ai2

Presenting work with



Weijia Shi, Akshita Bhagia, Kevin Farhat

+ Other collaborators at Ai2!

What This Talk Covers

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

In this talk, we'll show you we can use MoE to enable **training on distributed datasets owned by different parties**

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

In this talk, we'll show you we can use MoE to enable **training on distributed datasets owned by different parties**

- Why is important?

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

In this talk, we'll show you we can use MoE to enable **training on distributed datasets owned by different parties**

- Why is important?
- What are other alternative approaches?

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

In this talk, we'll show you we can use MoE to enable **training on distributed datasets owned by different parties**

- Why is important?
- What are other alternative approaches?
- How do we address the challenges using MoE?

What This Talk Covers

Standard MoEs: mainly for training & inference efficiency

In this talk, we'll show you we can use MoE to enable **training on distributed datasets owned by different parties**

- Why is important?
- What are other alternative approaches?
- How do we address the challenges using MoE?
- What results do we get?

Standard LM Training

Centralized access to all data during training

Data access is binary: available or unavailable

In Reality

? Centralized access to all data during training

Data access is binary: available or unavailable

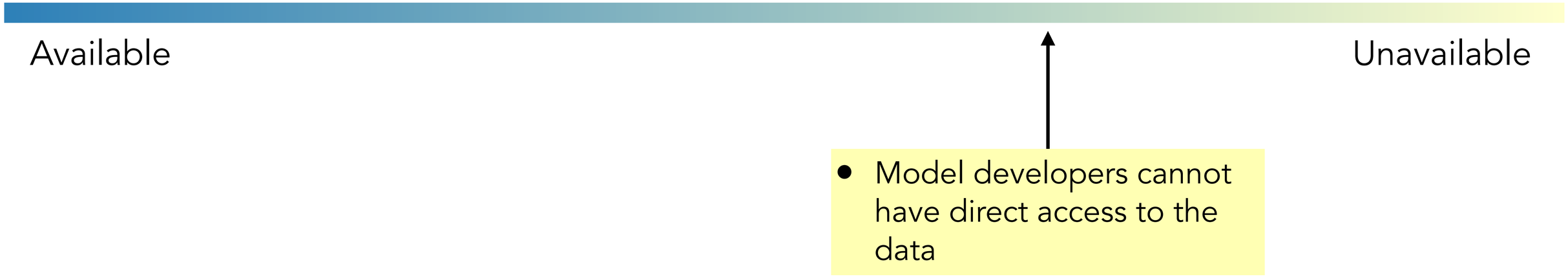
Available

Unavailable

In Reality

? Centralized access to all data during training

Data access is binary: available or unavailable



- Model developers cannot have direct access to the data

In Reality

? Centralized access to all data during training

~~Data access is binary: available or unavailable~~

Available

Unavailable

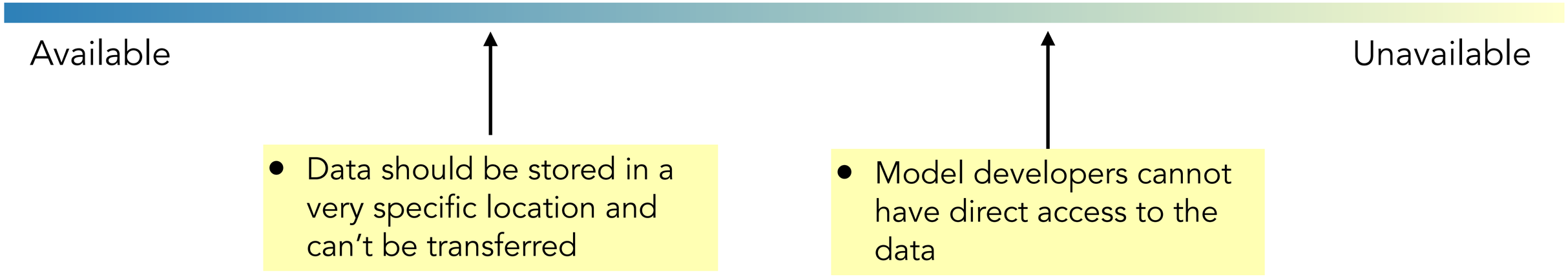
- Data should be stored in a very specific location and can’t be transferred

- Model developers cannot have direct access to the data

In Reality

? Centralized access to all data during training

~~Data access is binary: available or unavailable~~

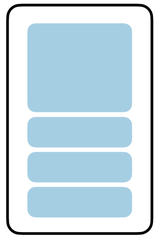


- Data should be stored in a very specific location and can't be transferred
- Model developers cannot have direct access to the data

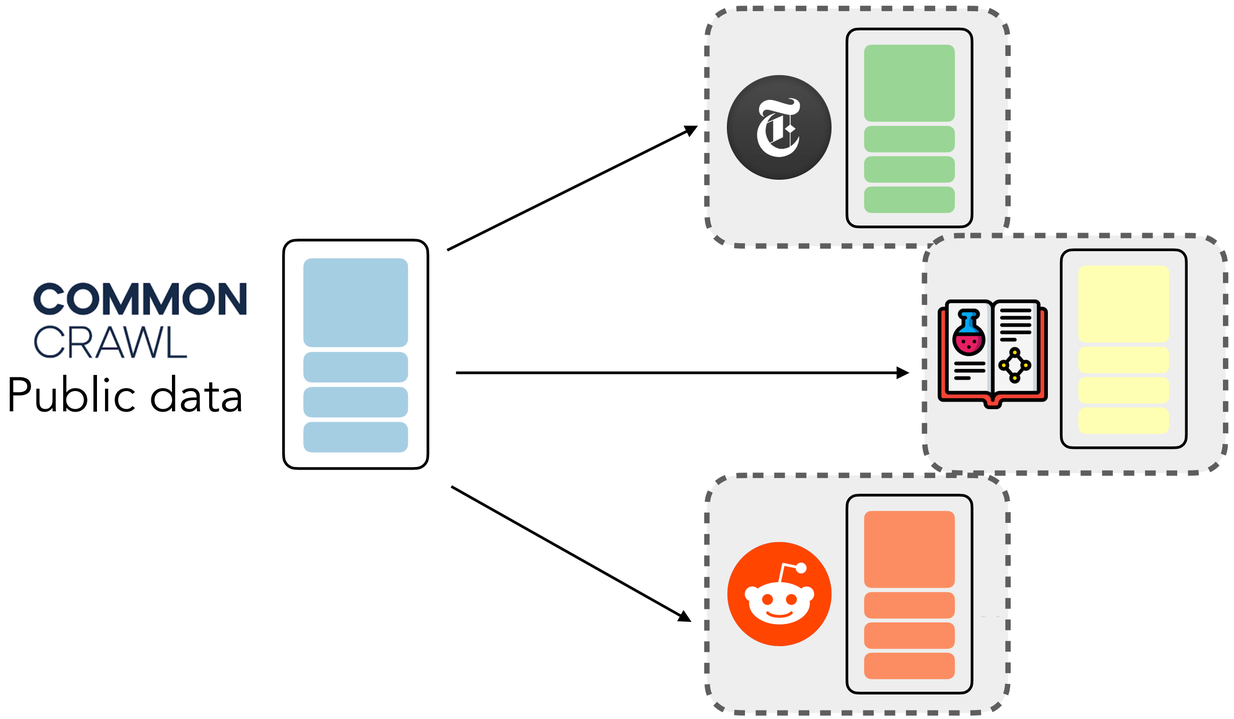
Other Restrictions: Data may become available at different times, may have expiration dates or owner might require opt-out.

FlexOlmo: Modular, Distributed Training

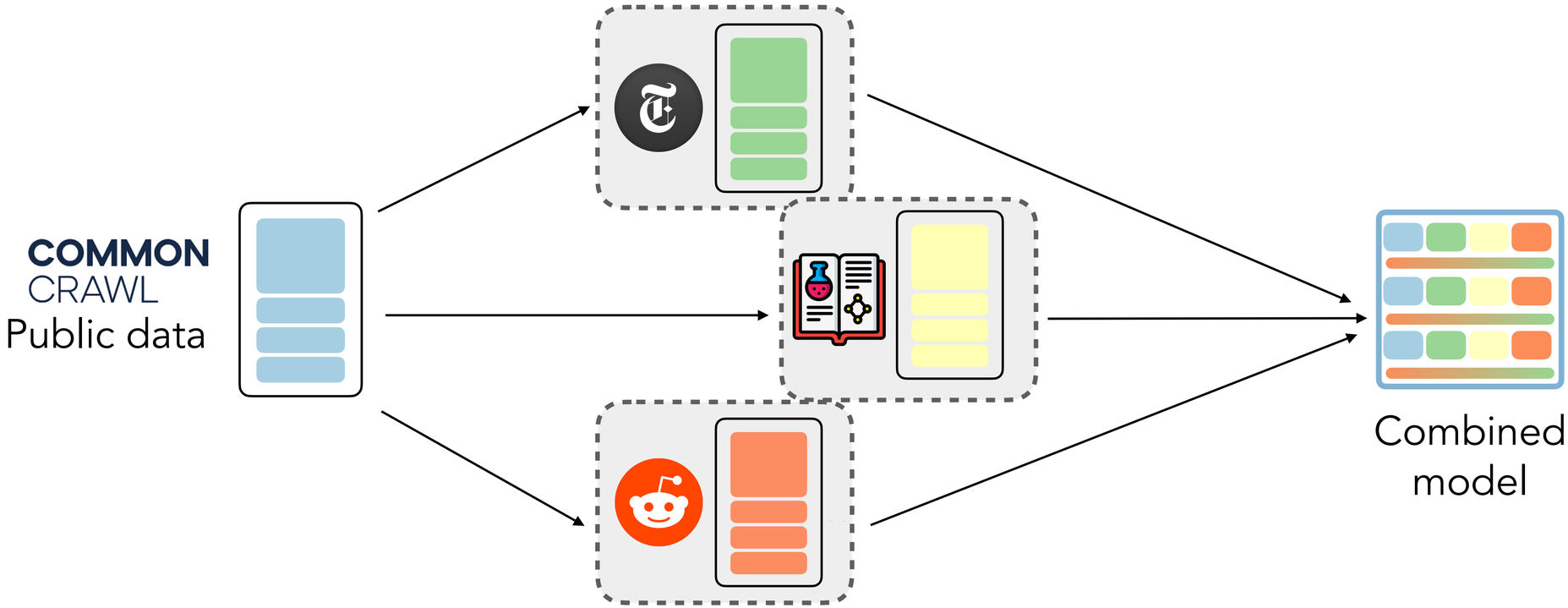
COMMON CRAWL
Public data



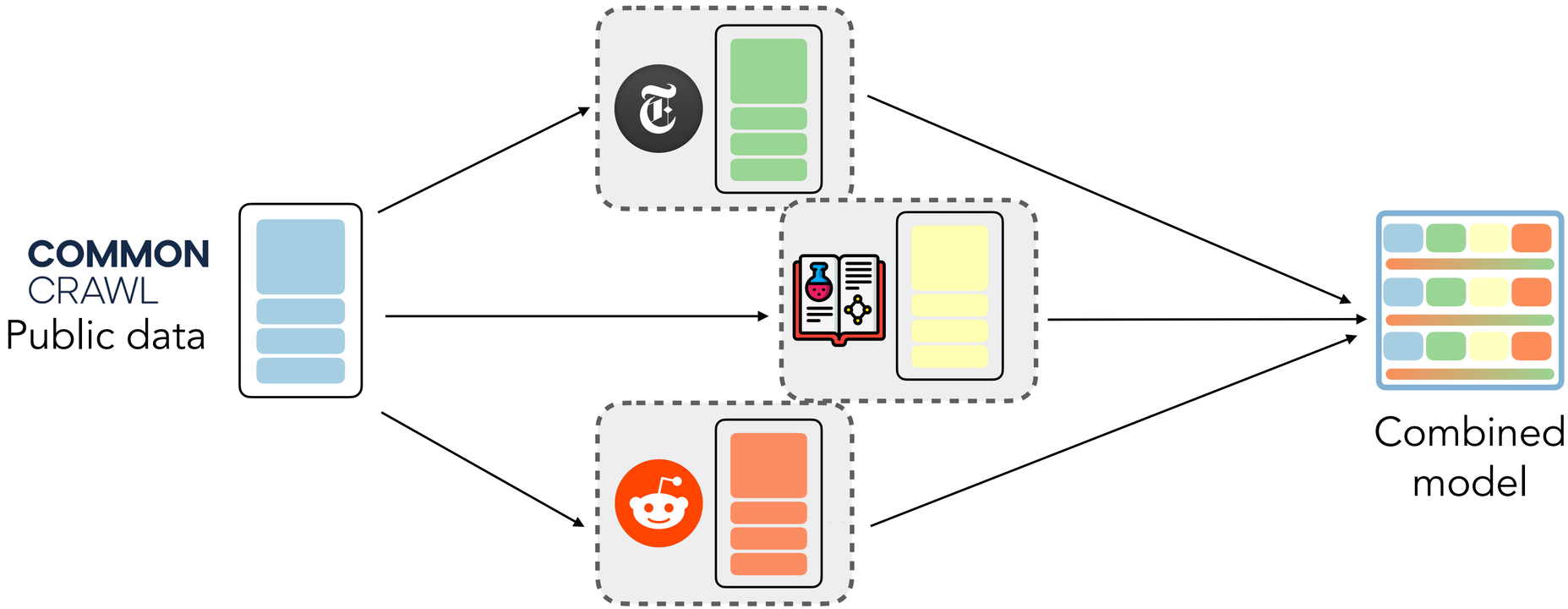
FlexOlmo: Modular, Distributed Training



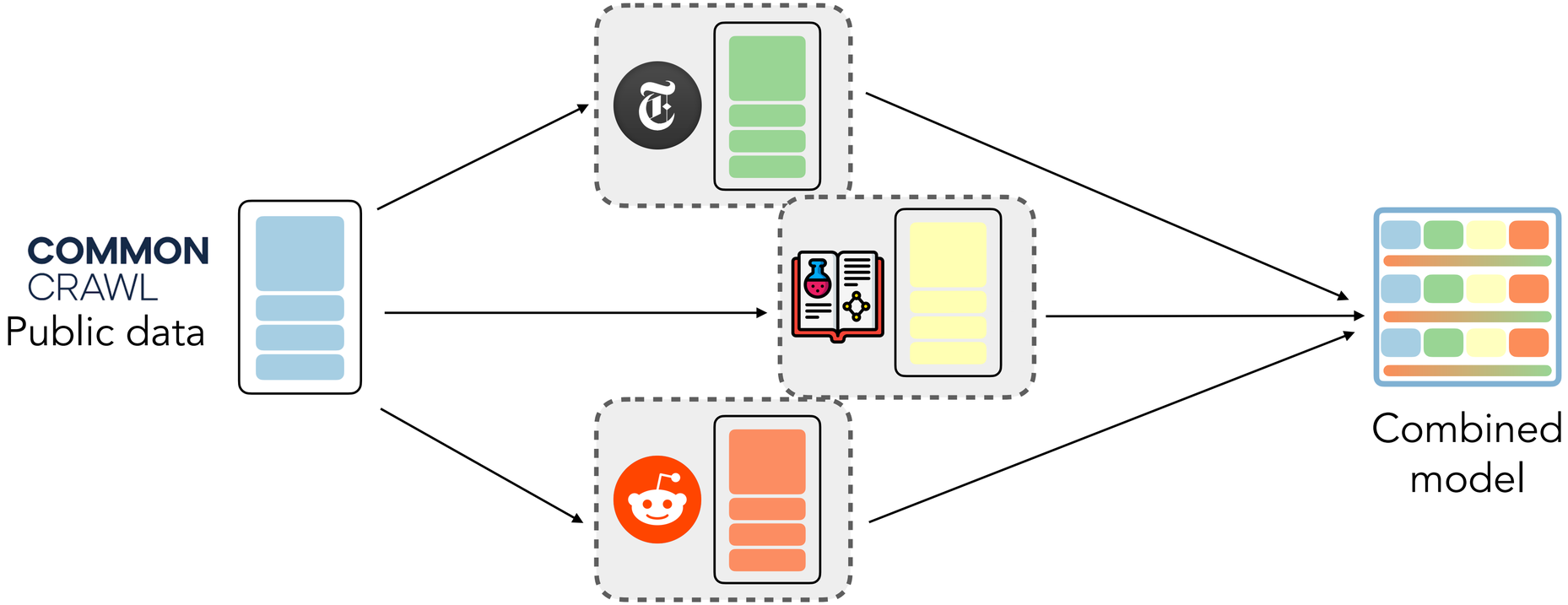
FlexOlmo: Modular, Distributed Training



FlexOlmo: Modular, Distributed Training

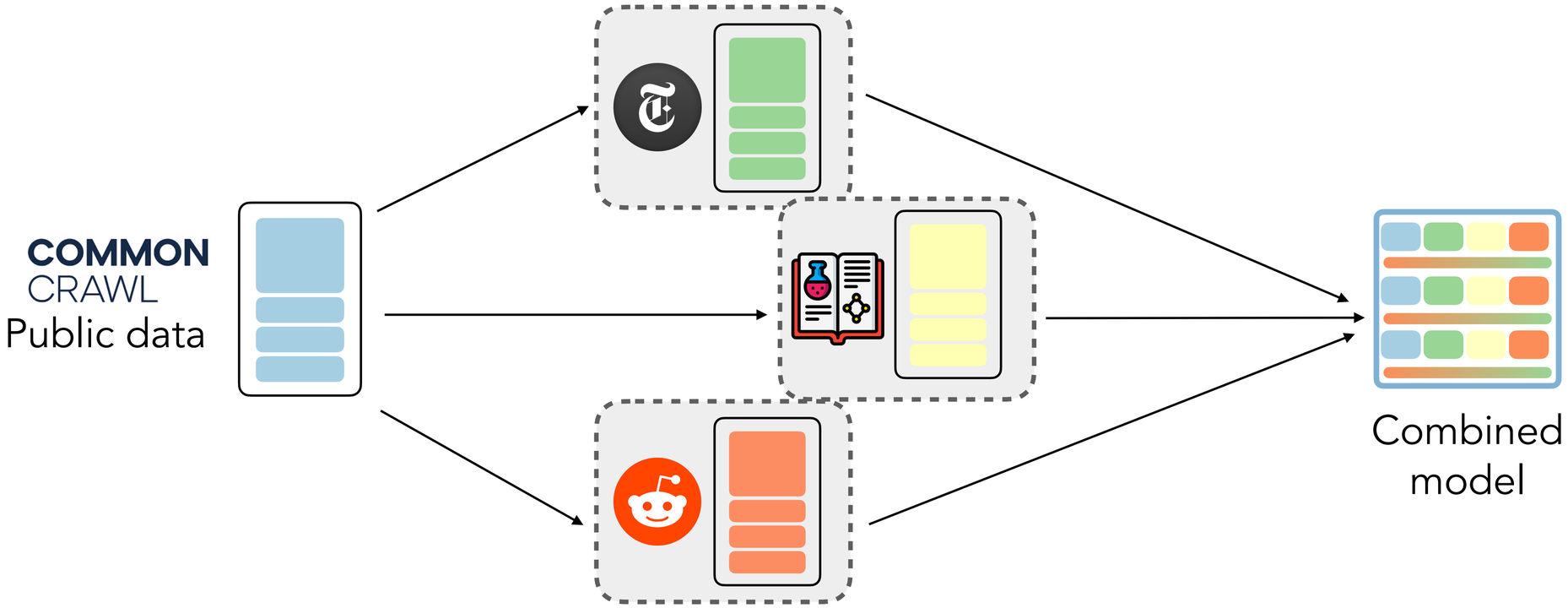


FlexOlmo: Modular, Distributed Training



Enables
data contribution
without data sharing

FlexOlmo: Modular, Distributed Training

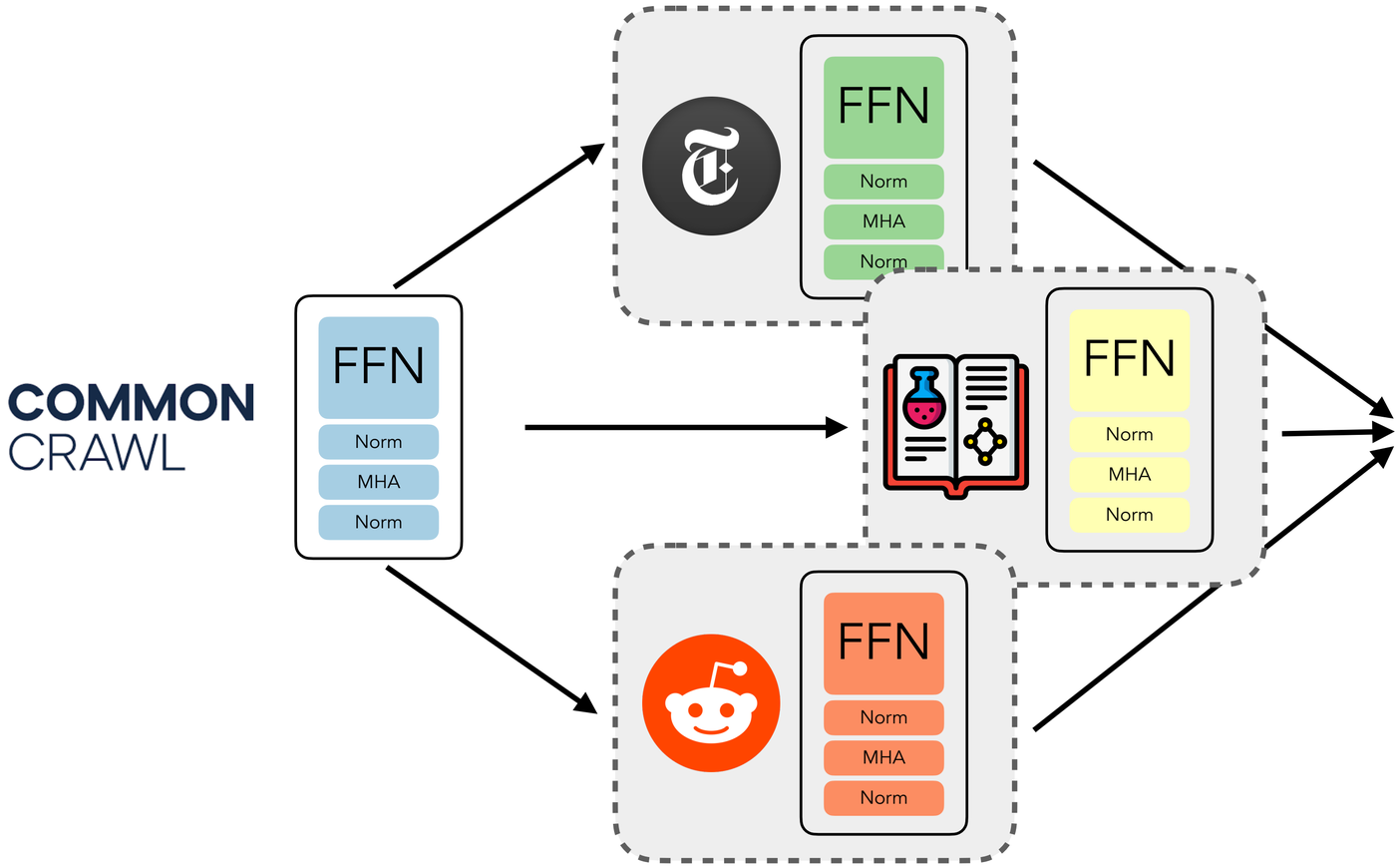


Enables
data contribution without data sharing

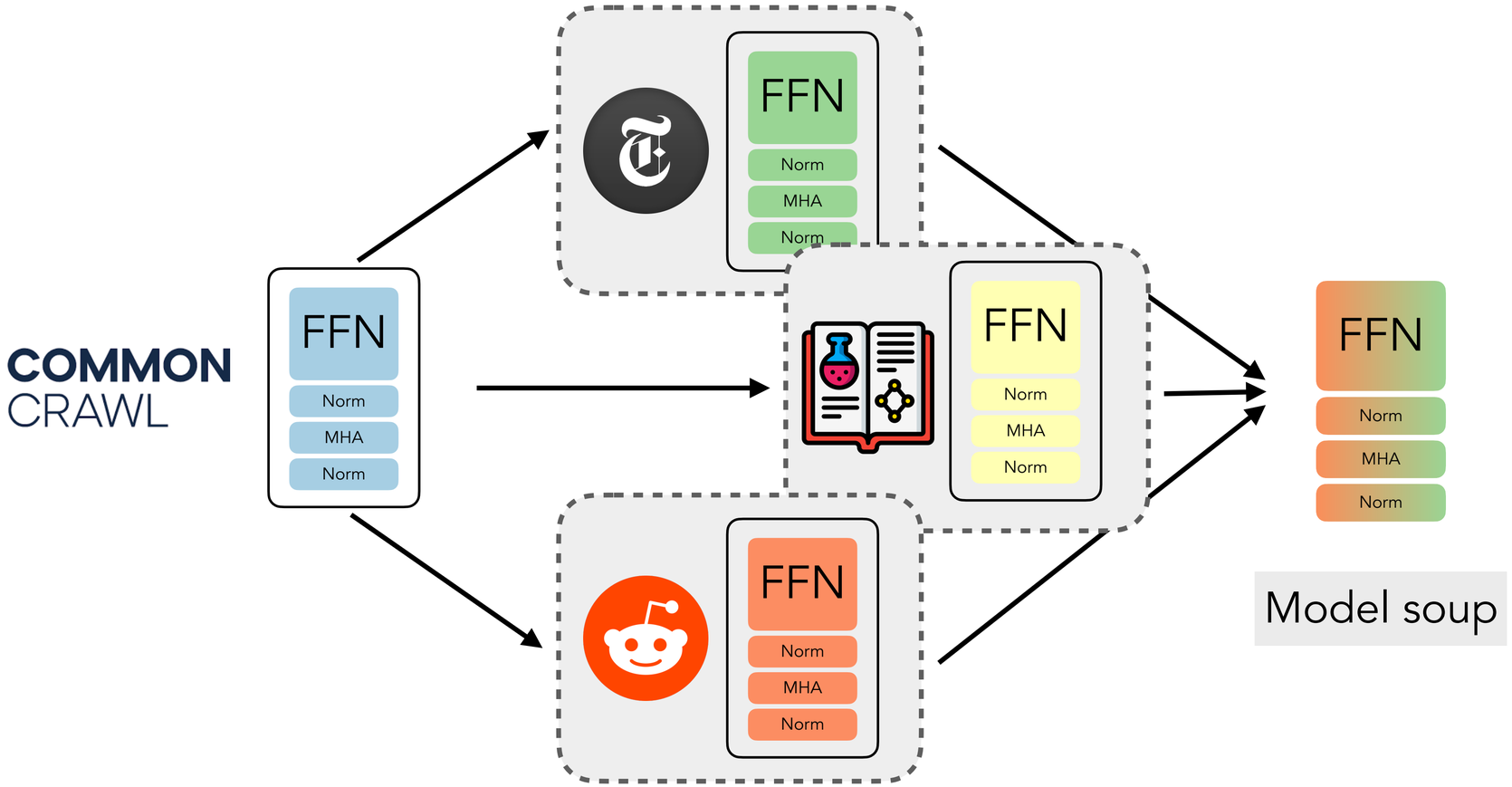
Supports **easy data addition/removal** with no further training

Option 1: Model merging

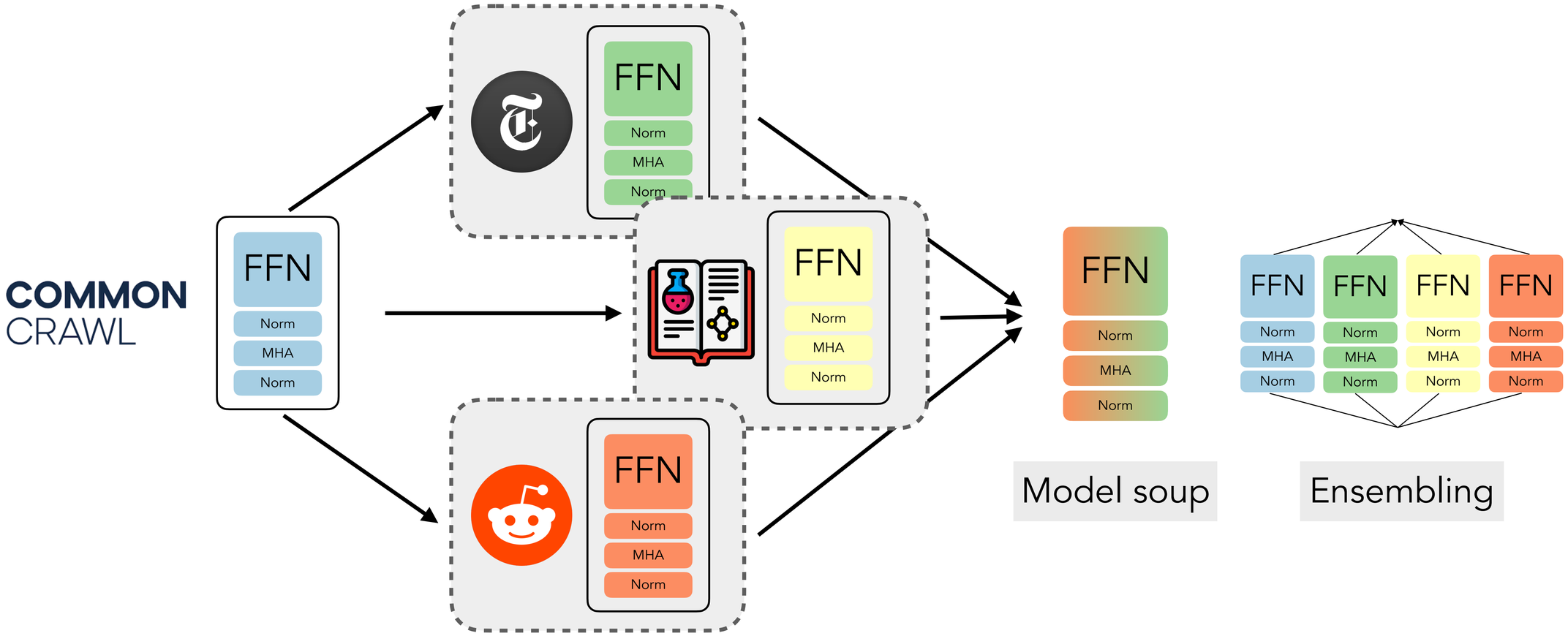
Option 1: Model merging



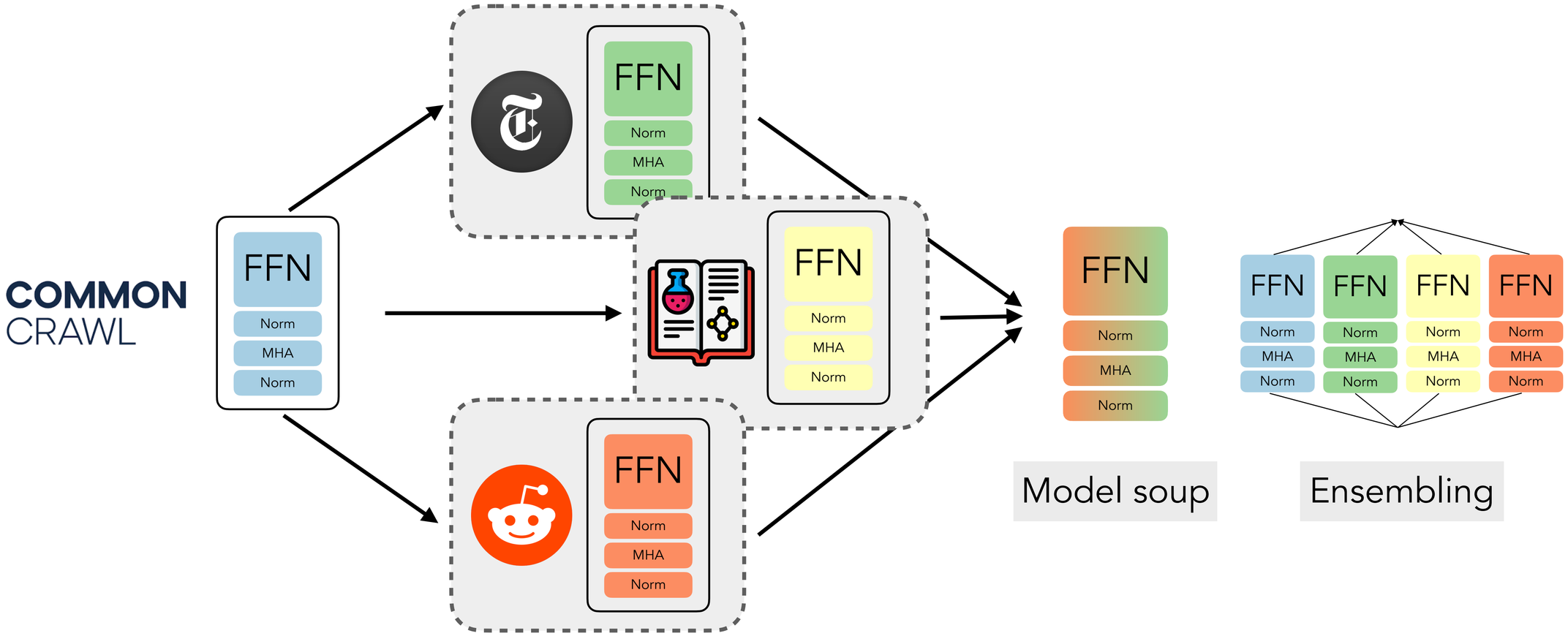
Option 1: Model merging



Option 1: Model merging



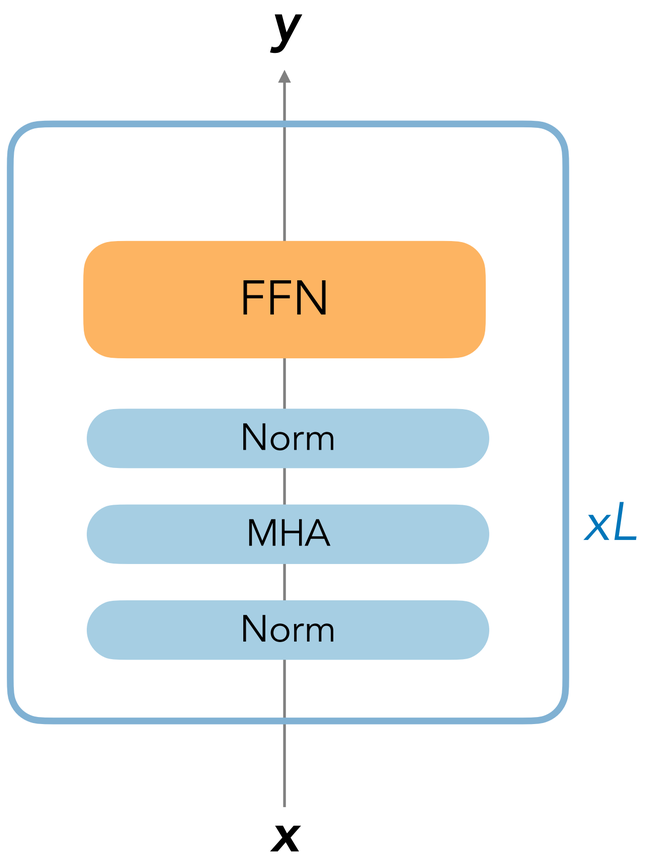
Option 1: Model merging



- Standard practice for boosting model performance
- However, merging models trained on disjoint datasets is harder

Option 2: MoE merging

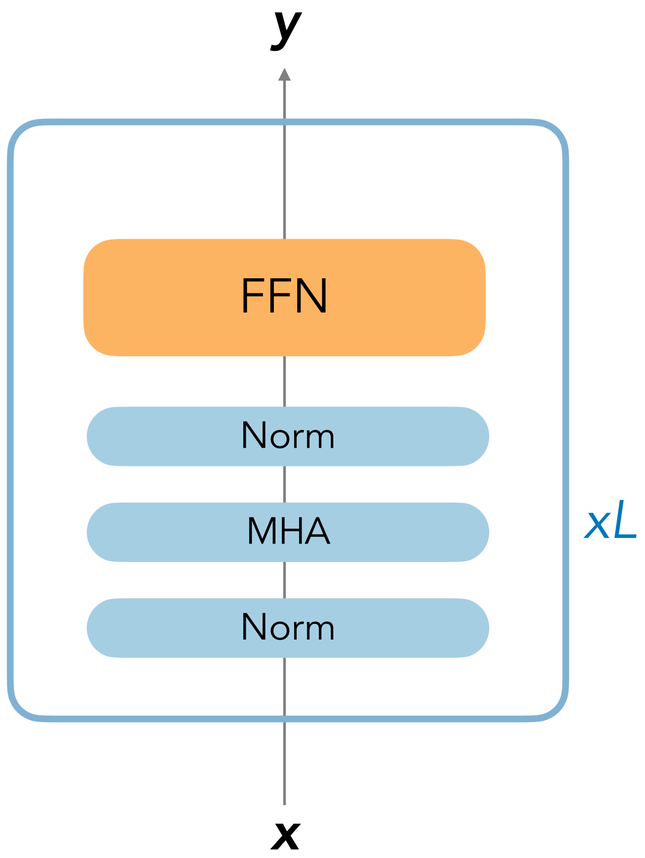
Option 2: MoE merging — What is MoE?

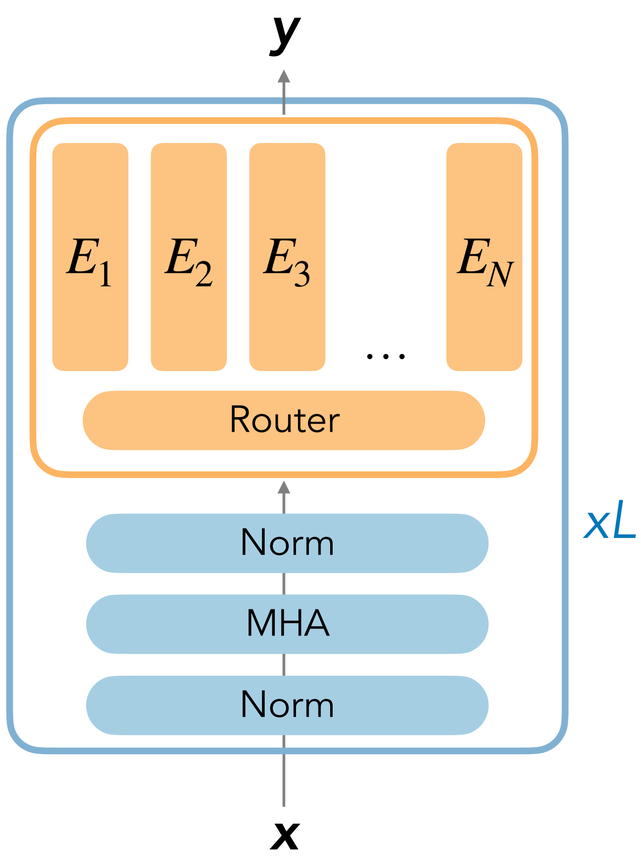


xL

*Residual connection not depicted

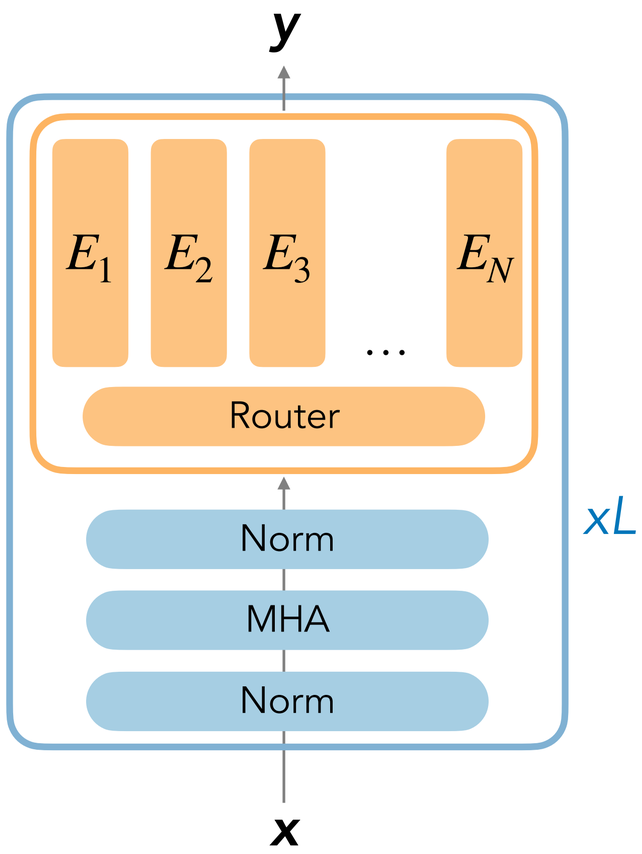
Option 2: MoE merging — What is MoE?





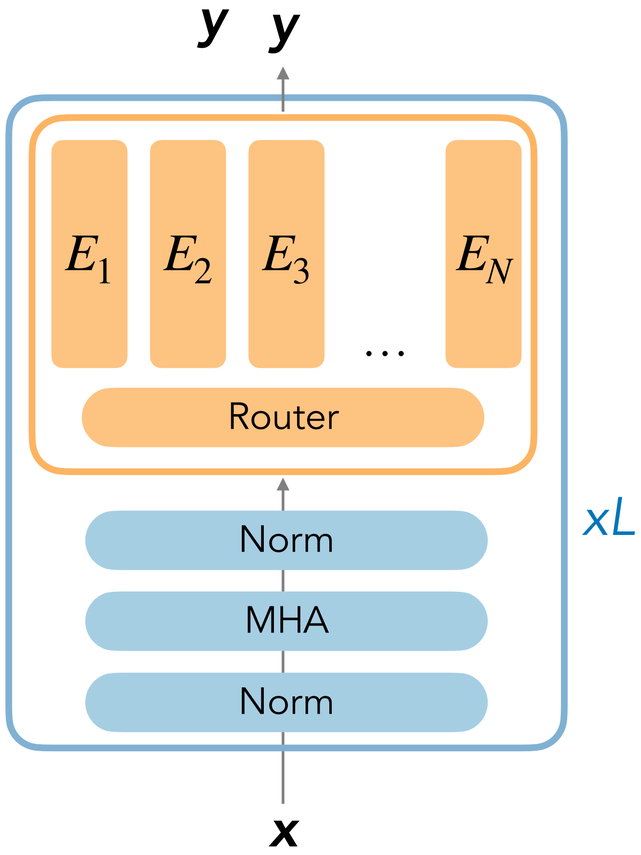
*Residual connection not depicted

Option 2: MoE merging — What is MoE?



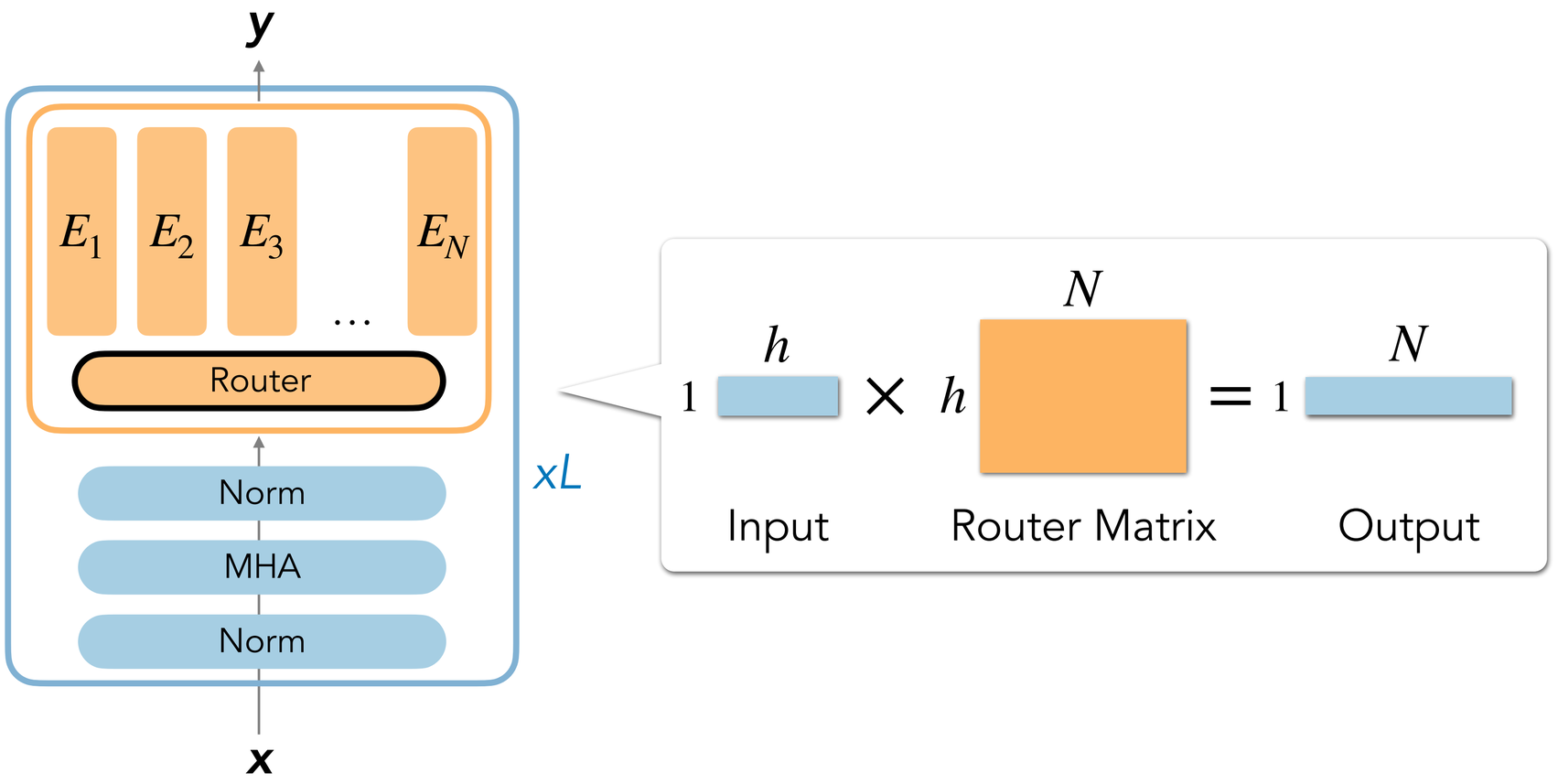
*Residual connection not depicted

Option 2: MoE merging — What is MoE?

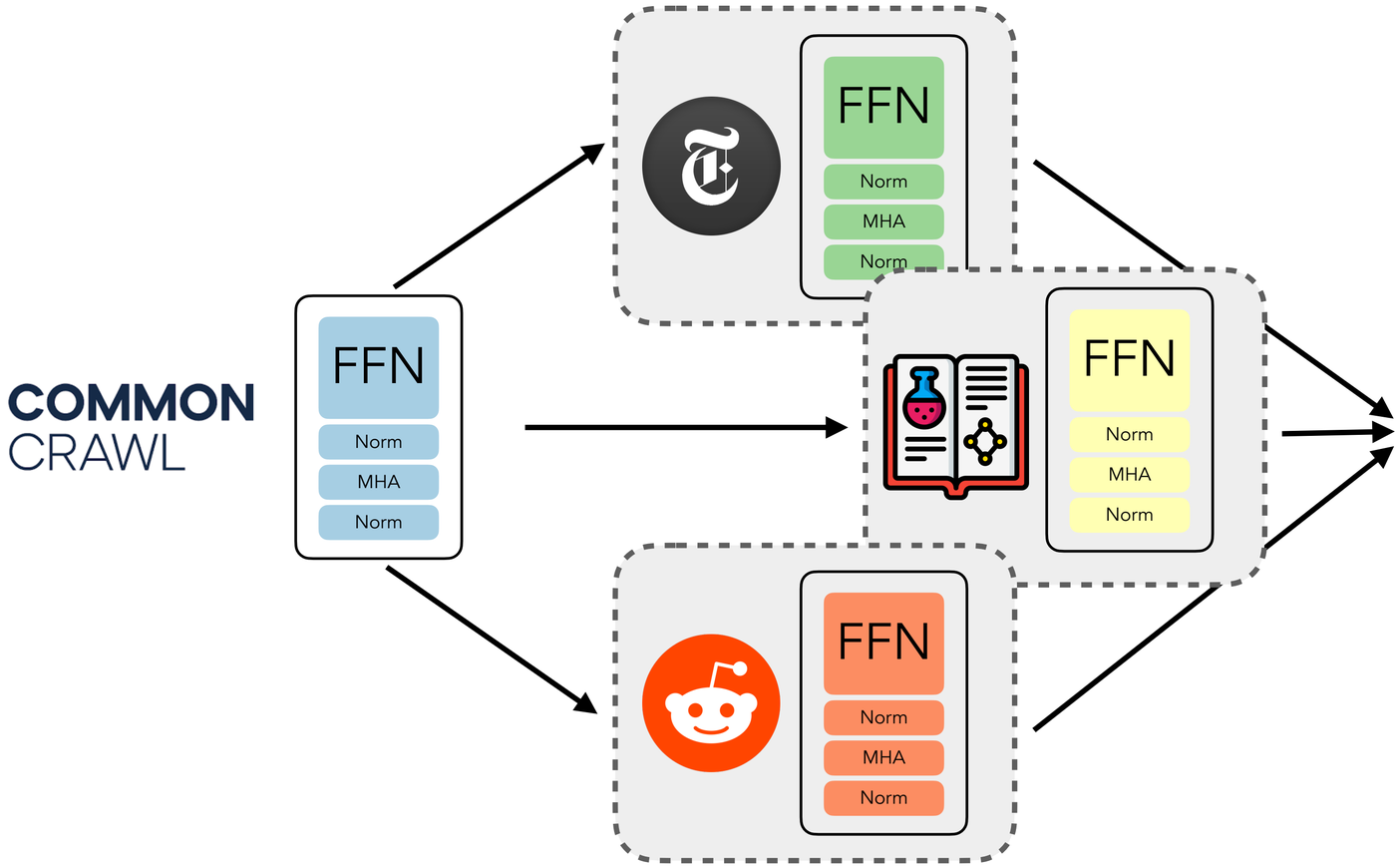


*Residual connection not depicted

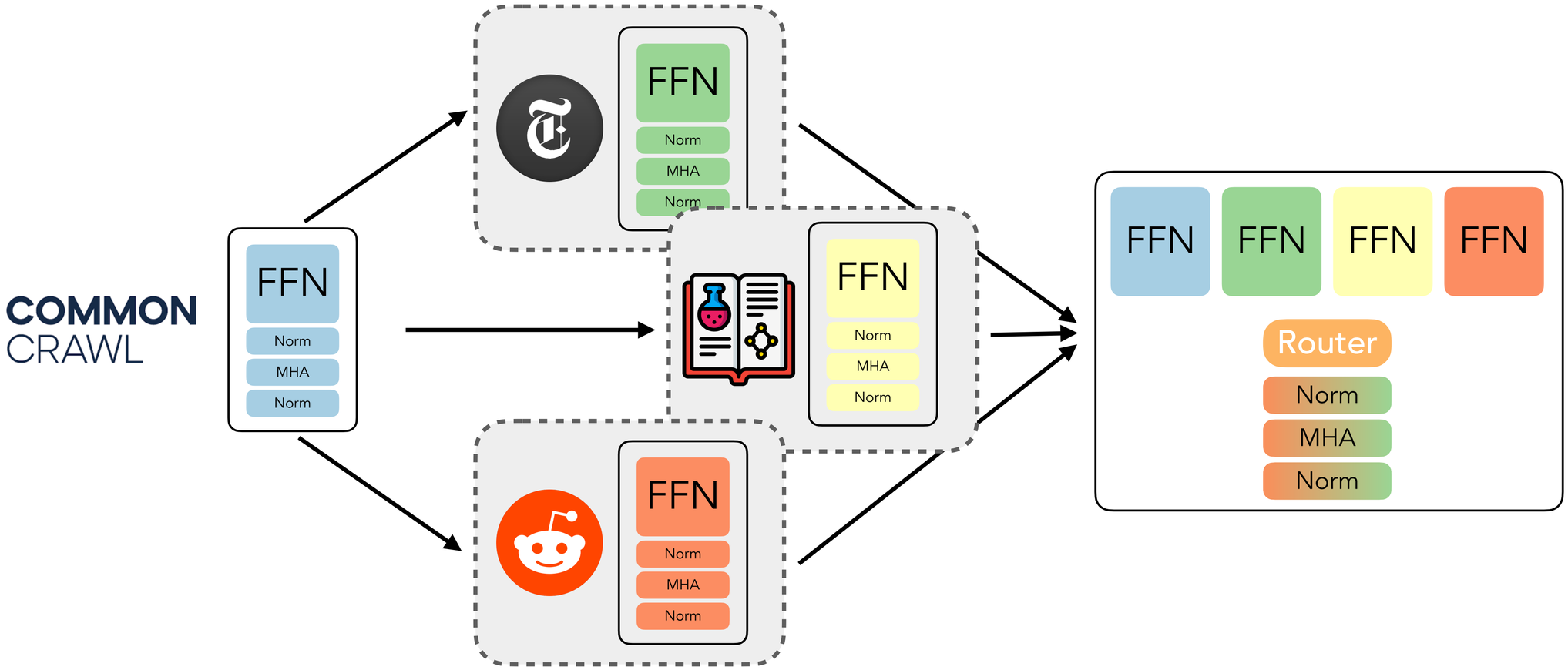
Option 2: MoE merging — What is MoE?



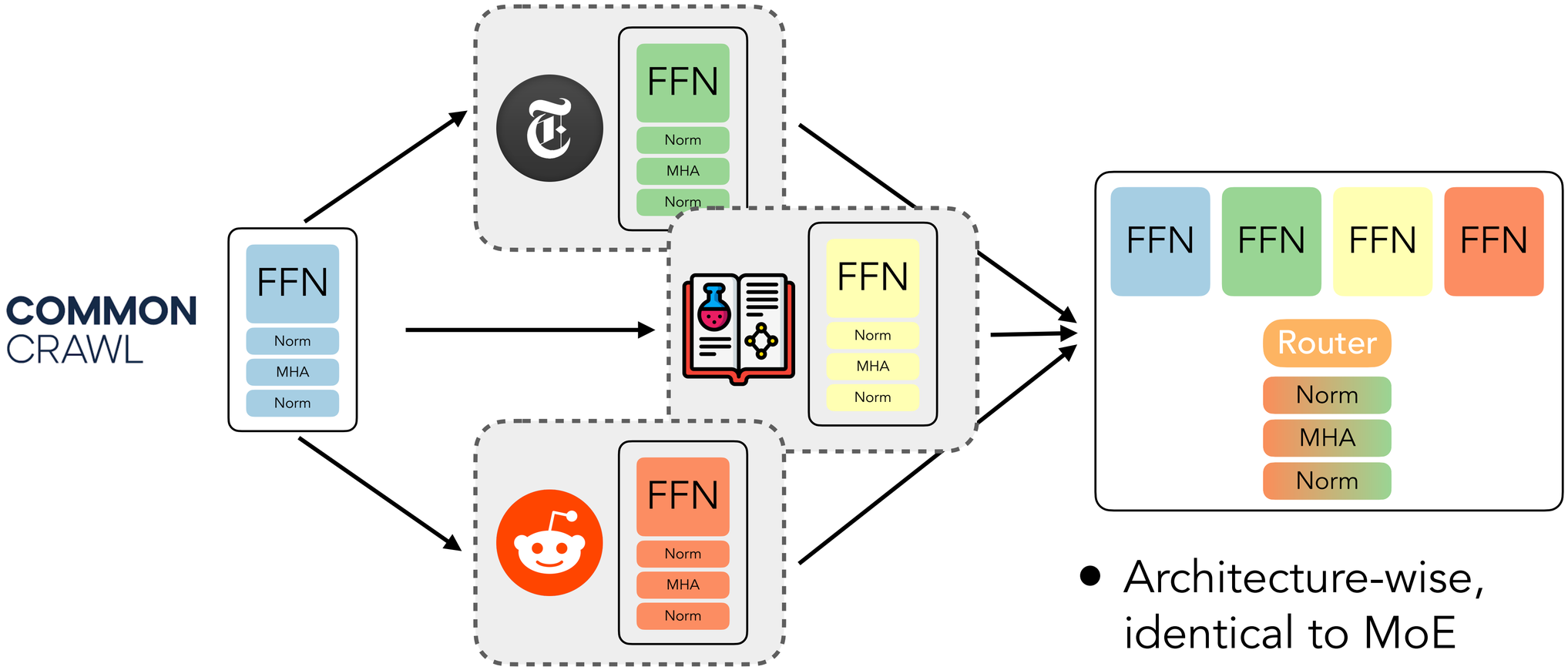
Option 2: MoE merging



Option 2: MoE merging

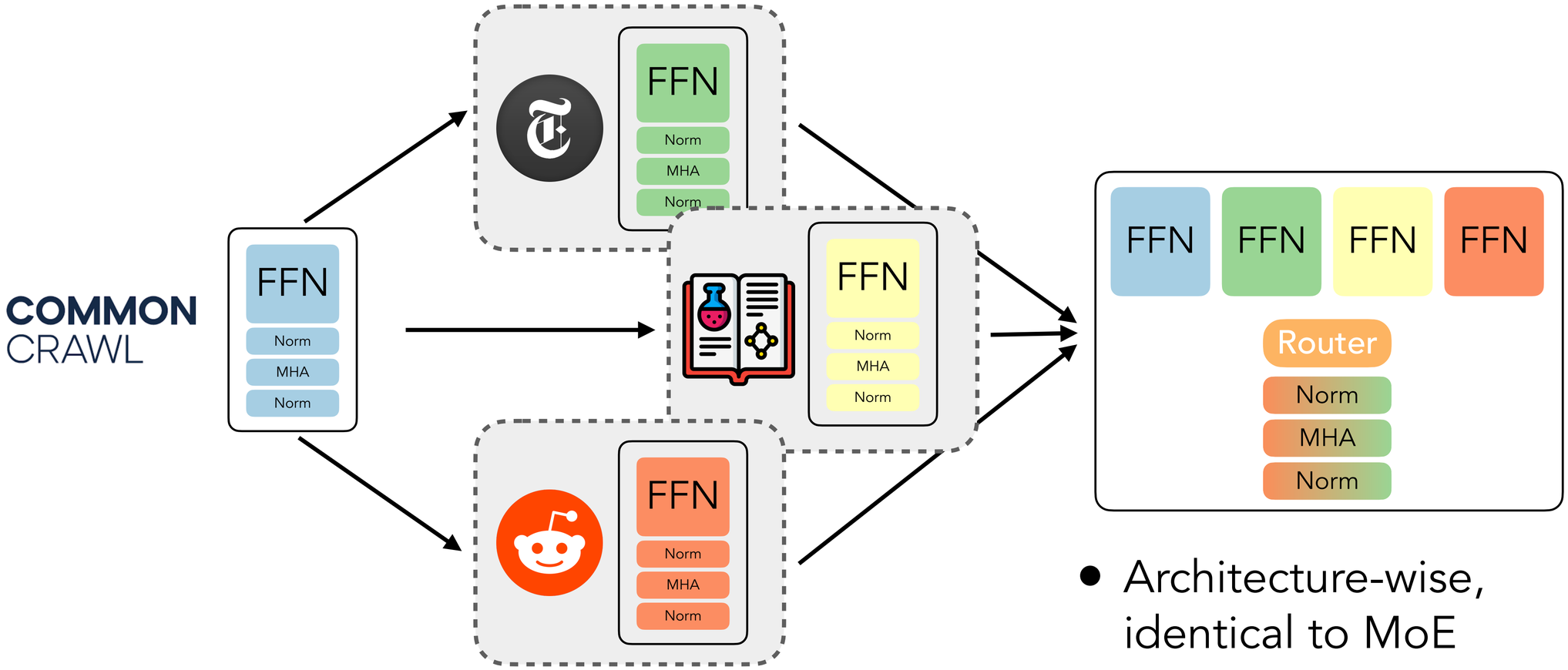


Option 2: MoE merging



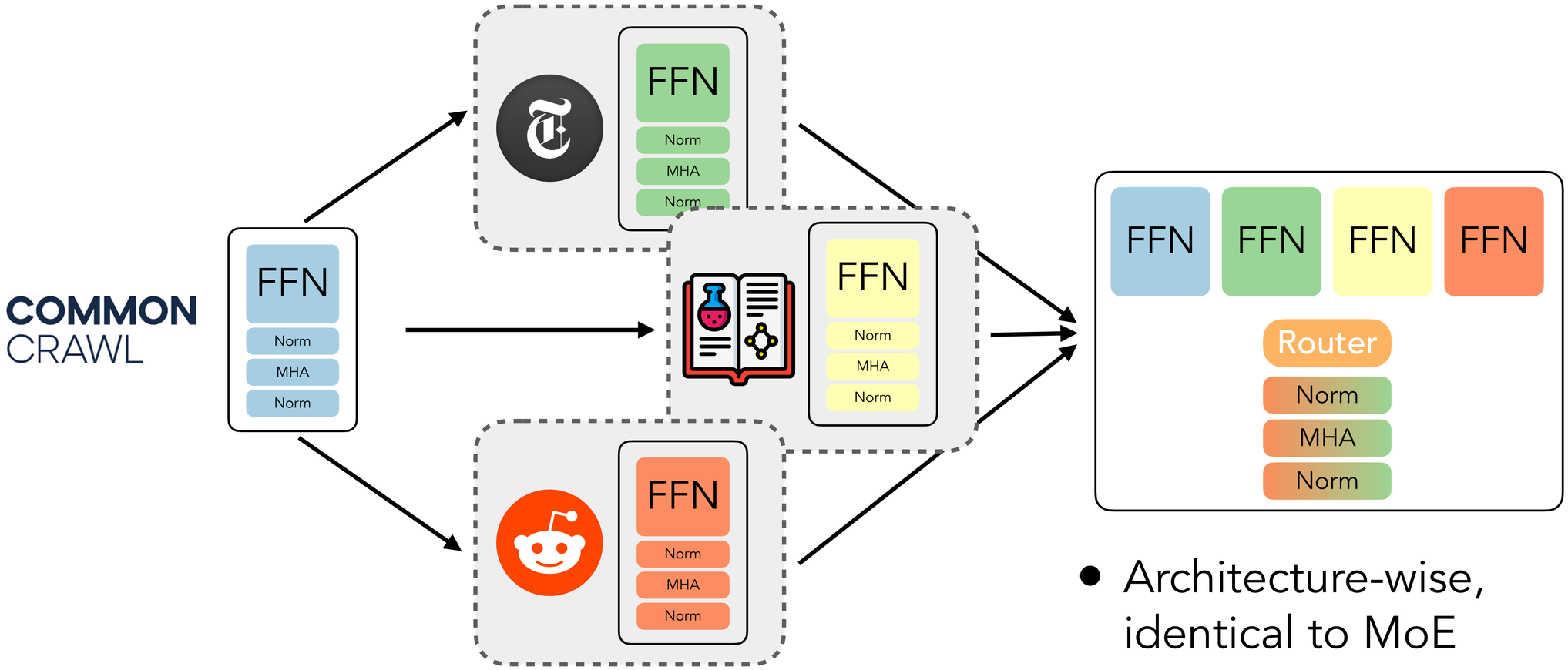
- Architecture-wise, identical to MoE

Option 2: MoE merging



- Architecture-wise, identical to MoE
- Requires training on all datasets after merging

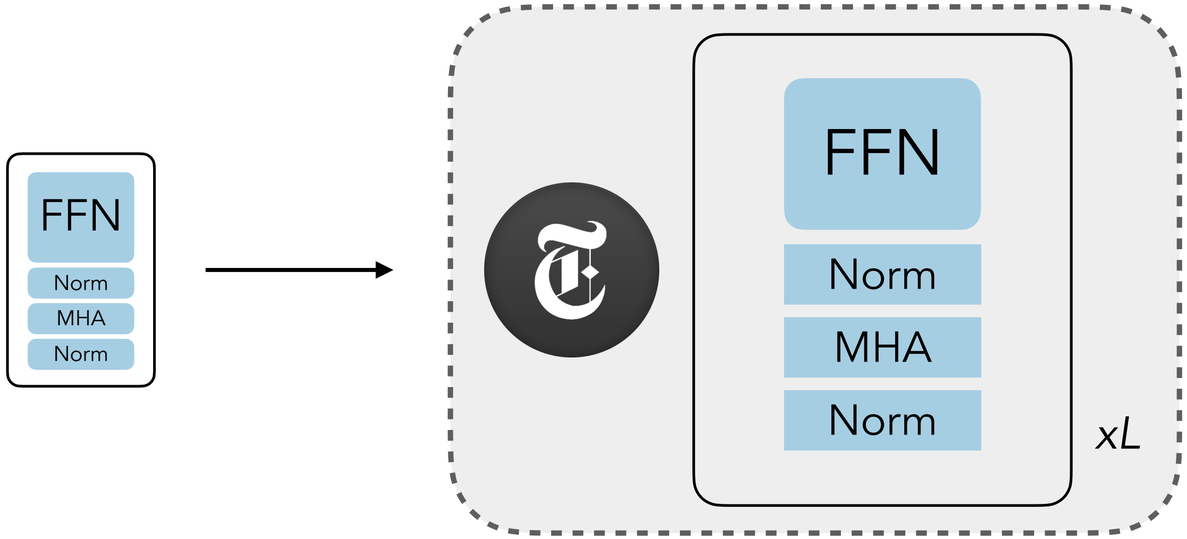
Option 2: MoE merging



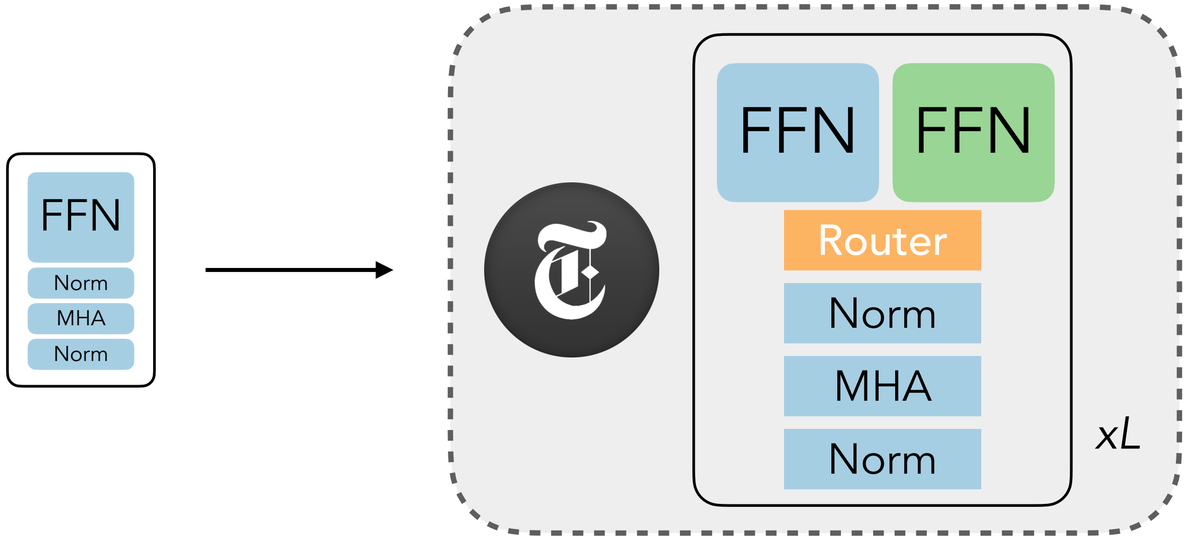
- Architecture-wise, identical to MoE
- Requires training on all after merging

Research Q: How do we leverage MoE but remove the needs for joint data access?

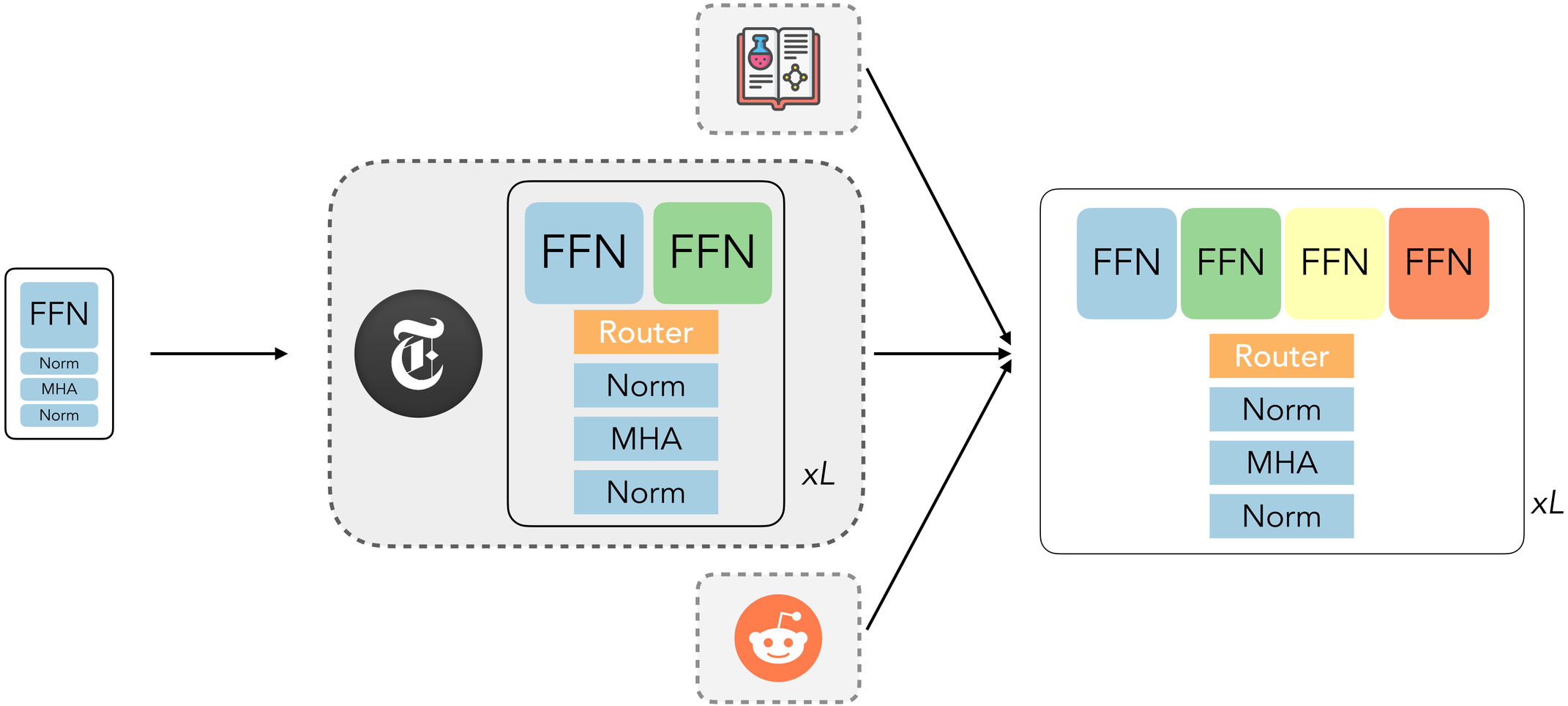
Solution 1: Learning to Coordinate



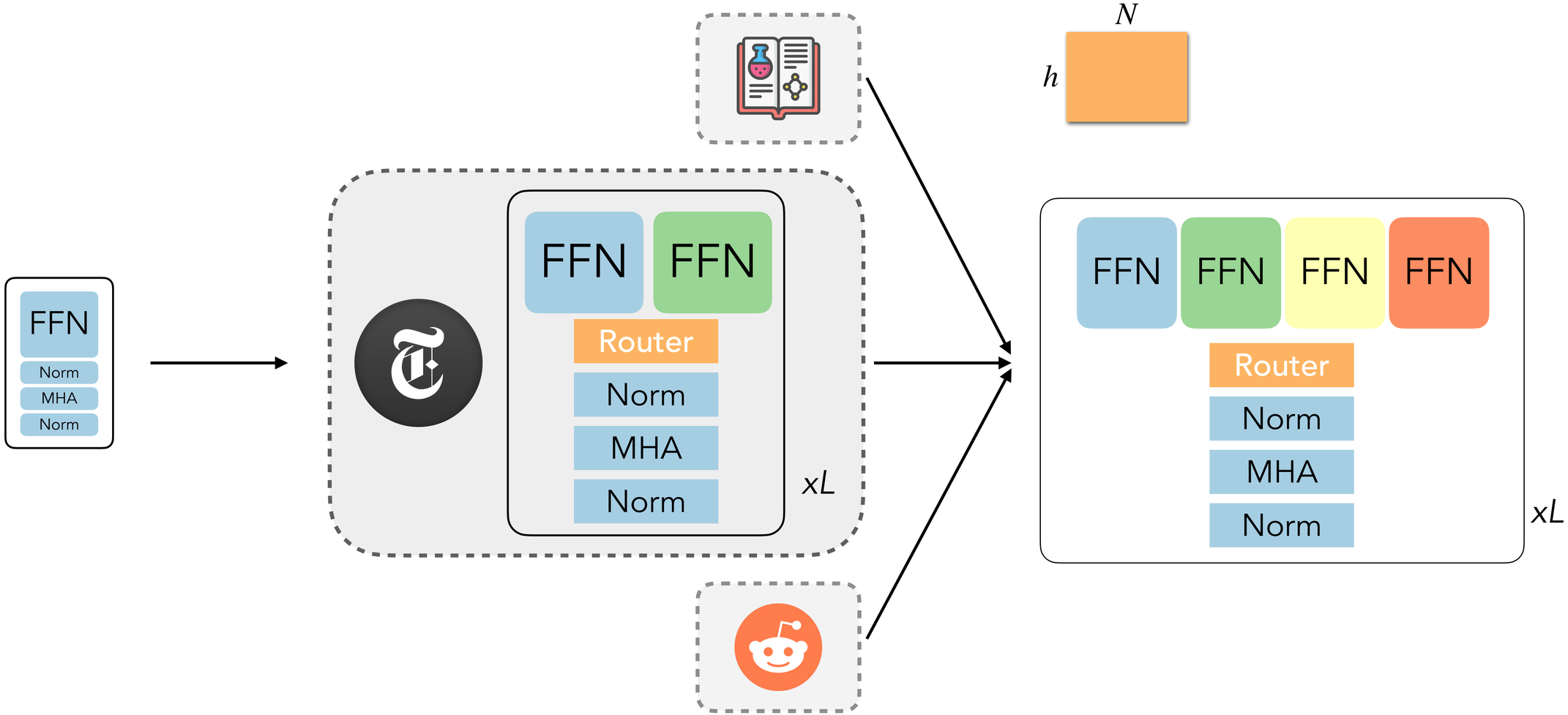
Solution 1: Learning to Coordinate



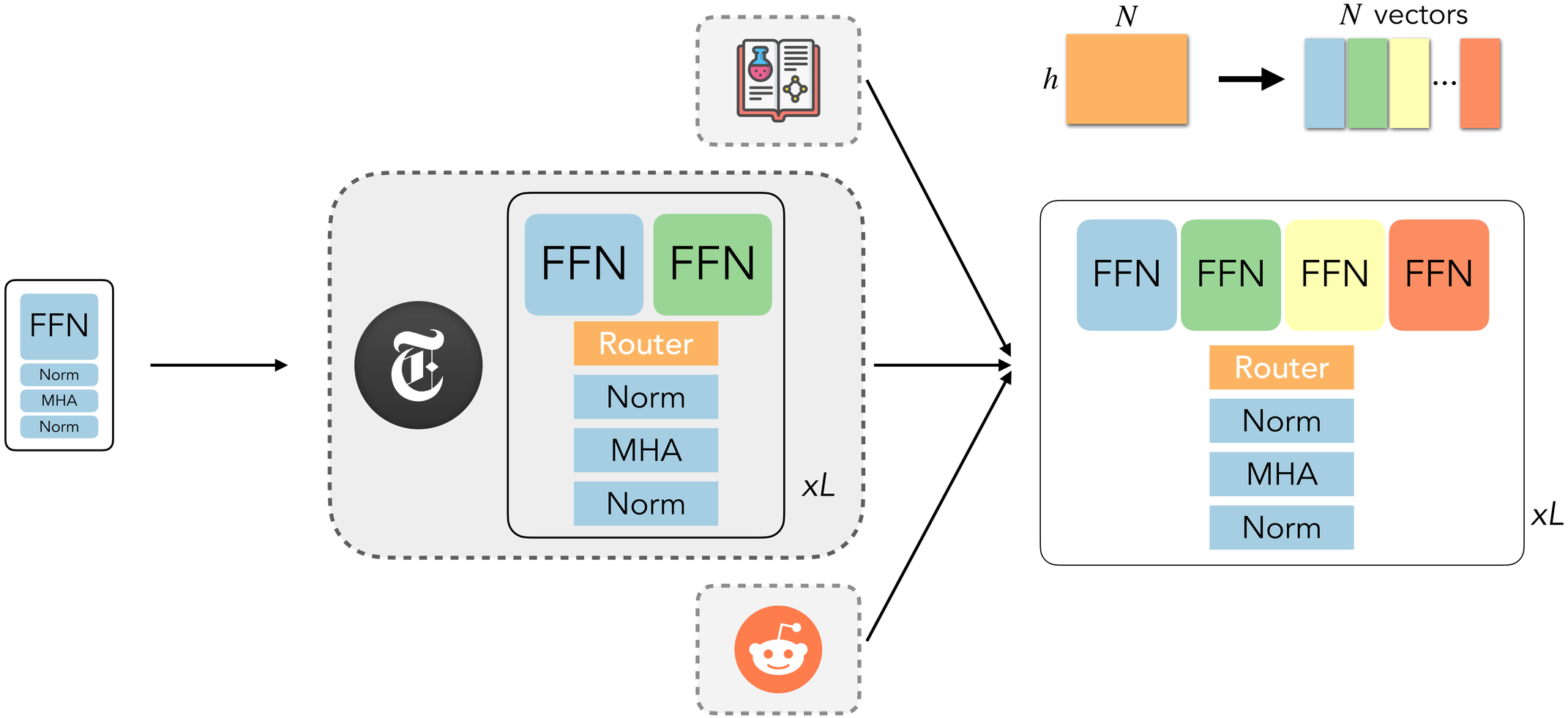
Solution 1: Learning to Coordinate



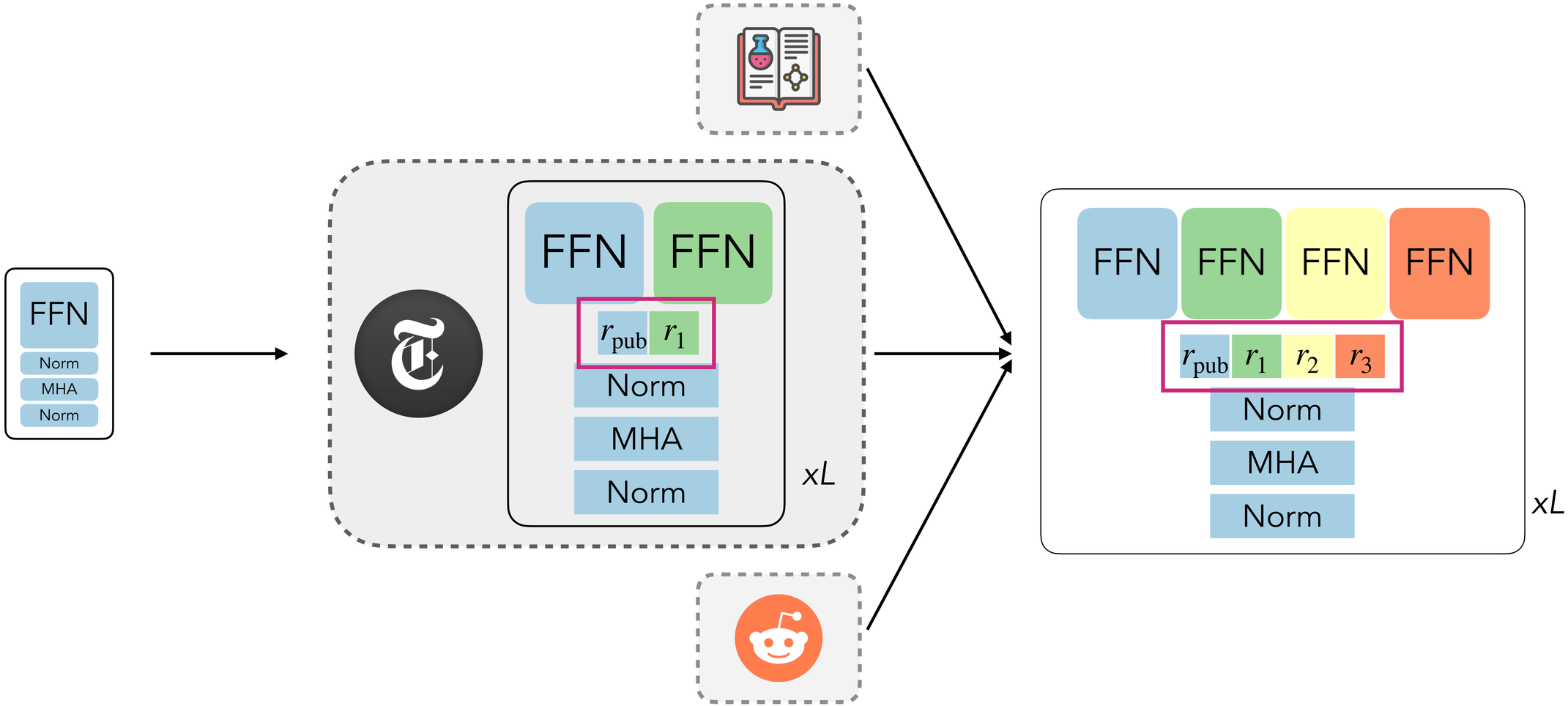
Solution 1: Learning to Coordinate



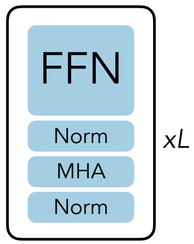
Solution 2: Nonparametric Router



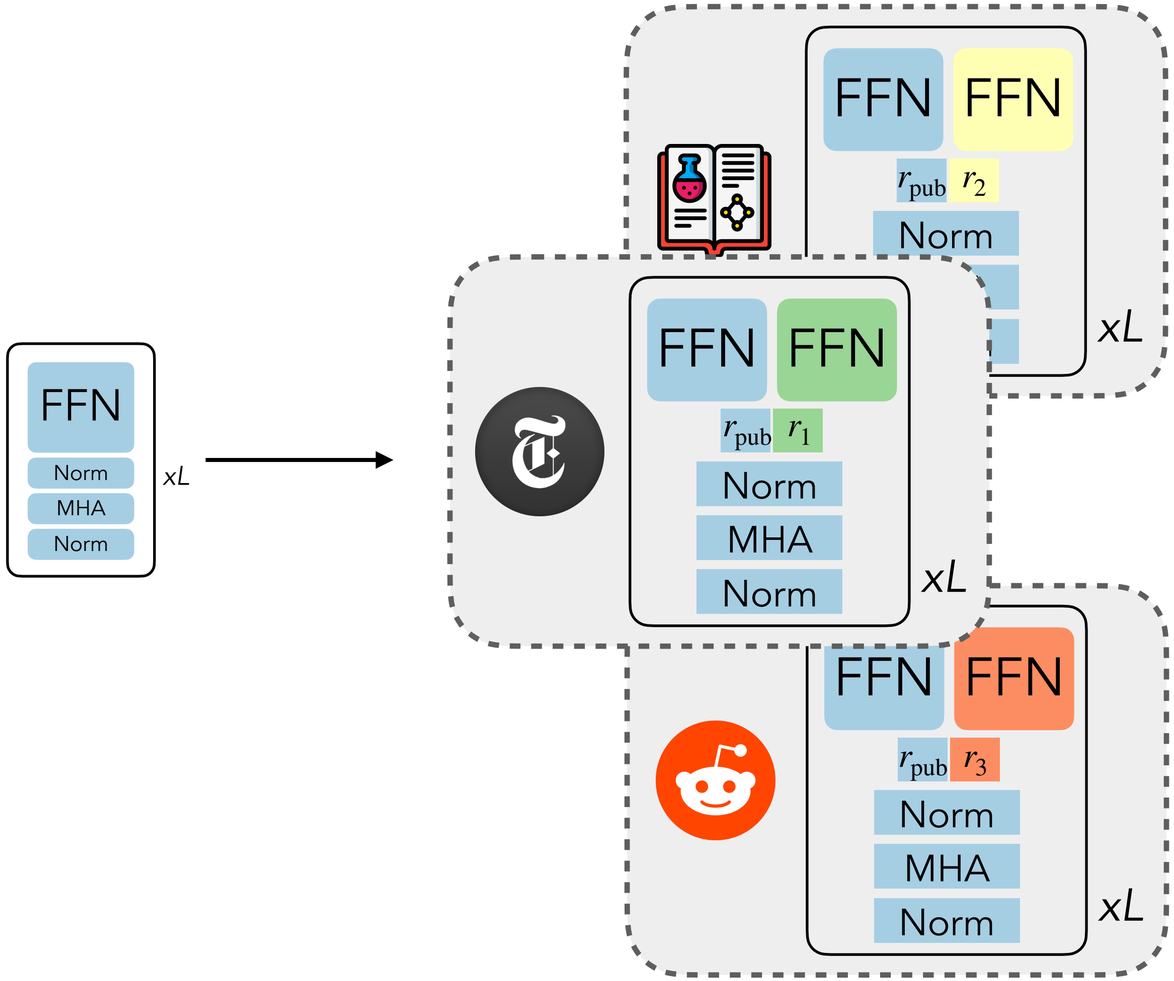
Solution 2: Nonparametric Router



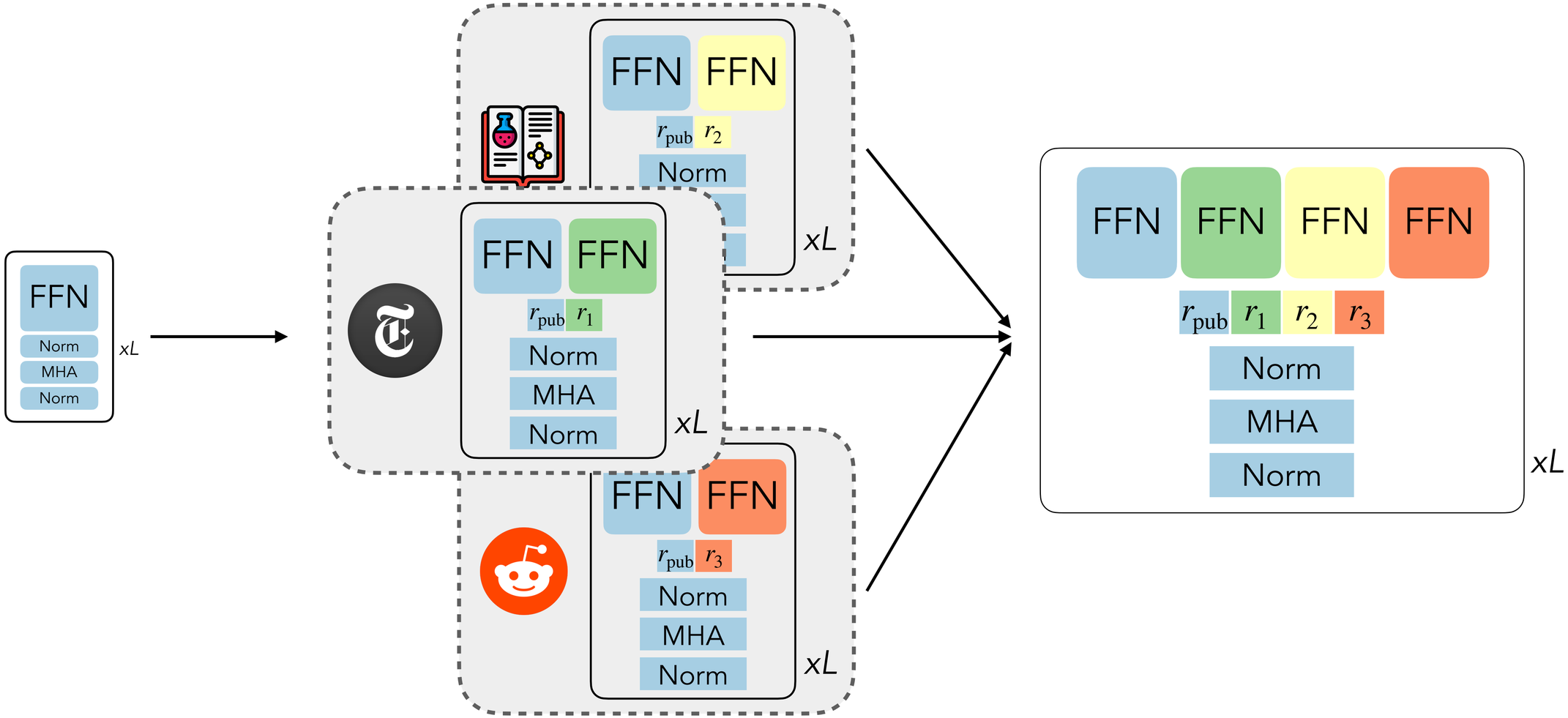
FlexOlmo: Summary



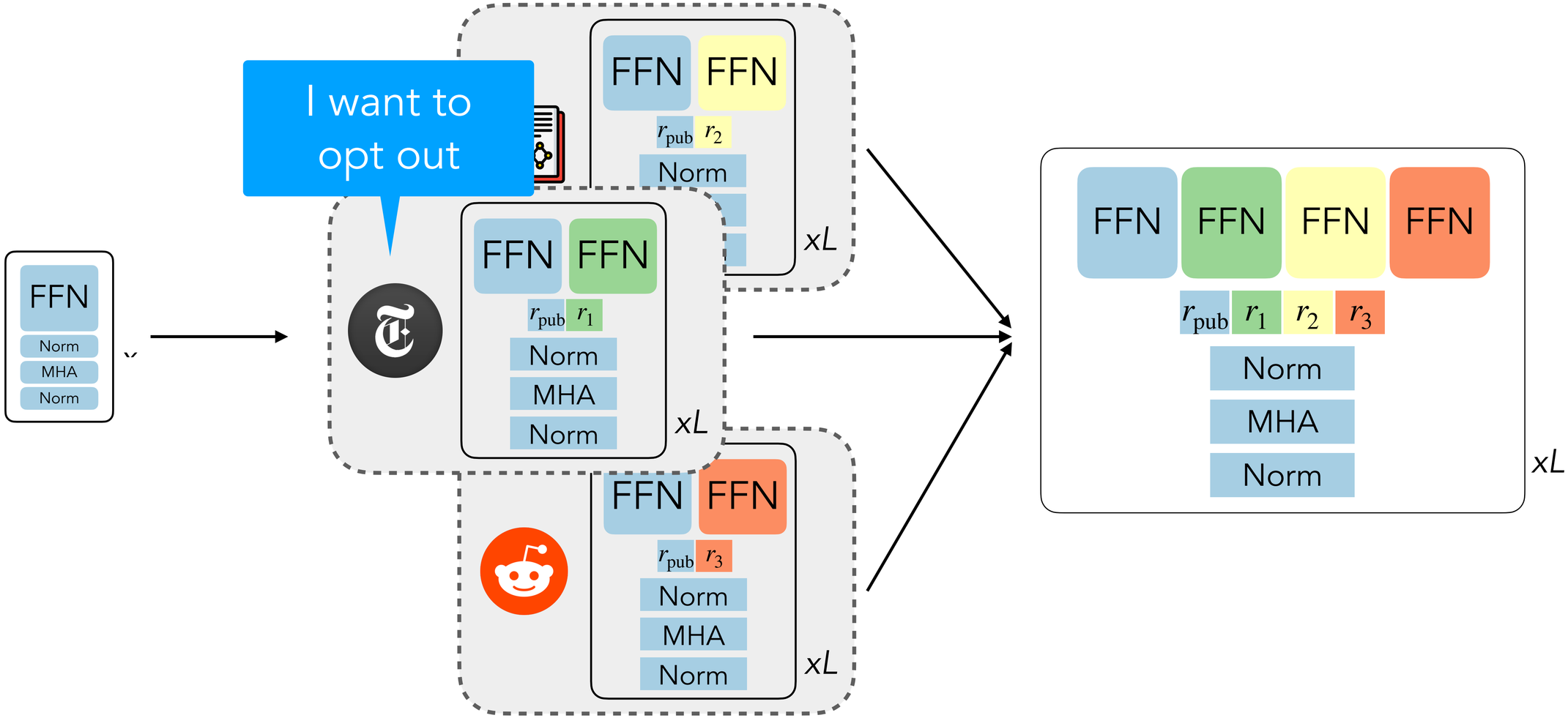
FlexOlmo: Summary



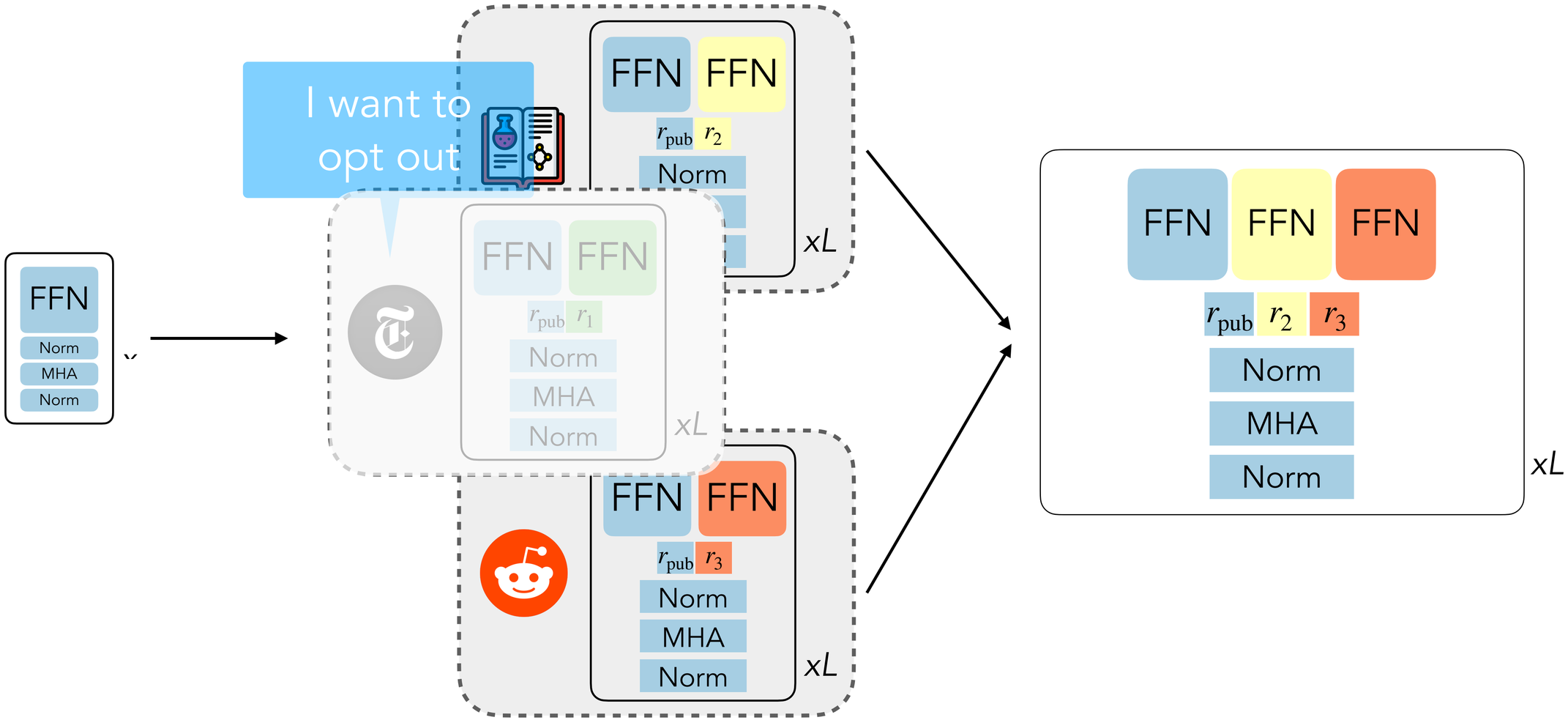
FlexOlmo: Summary



FlexOlmo: Summary



FlexOlmo: Summary



Experimental setup (1/2)

COMMON CRAWL

Public data

Experimental setup (1/2)

COMMON CRAWL

Public data

Math

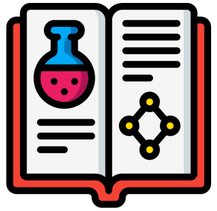
Code

Creative writings

Papers

News

Reddit



Educational Text

Experimental setup (1/2)

COMMON CRAWL

Public data

Math

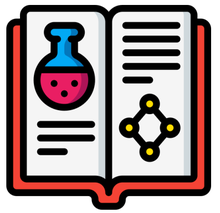
Code

Creative writings

Papers

News

Reddit



Educational Text

← Actual proprietary data (Can't publicly acquire as of now)

Experimental setup (1/2)

COMMON CRAWL

Public data

Math

Code

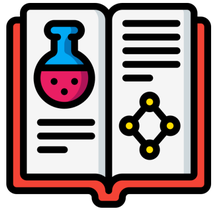
Creative writings

Papers

← Simulated proprietary data (Realistic domains)

News

Reddit



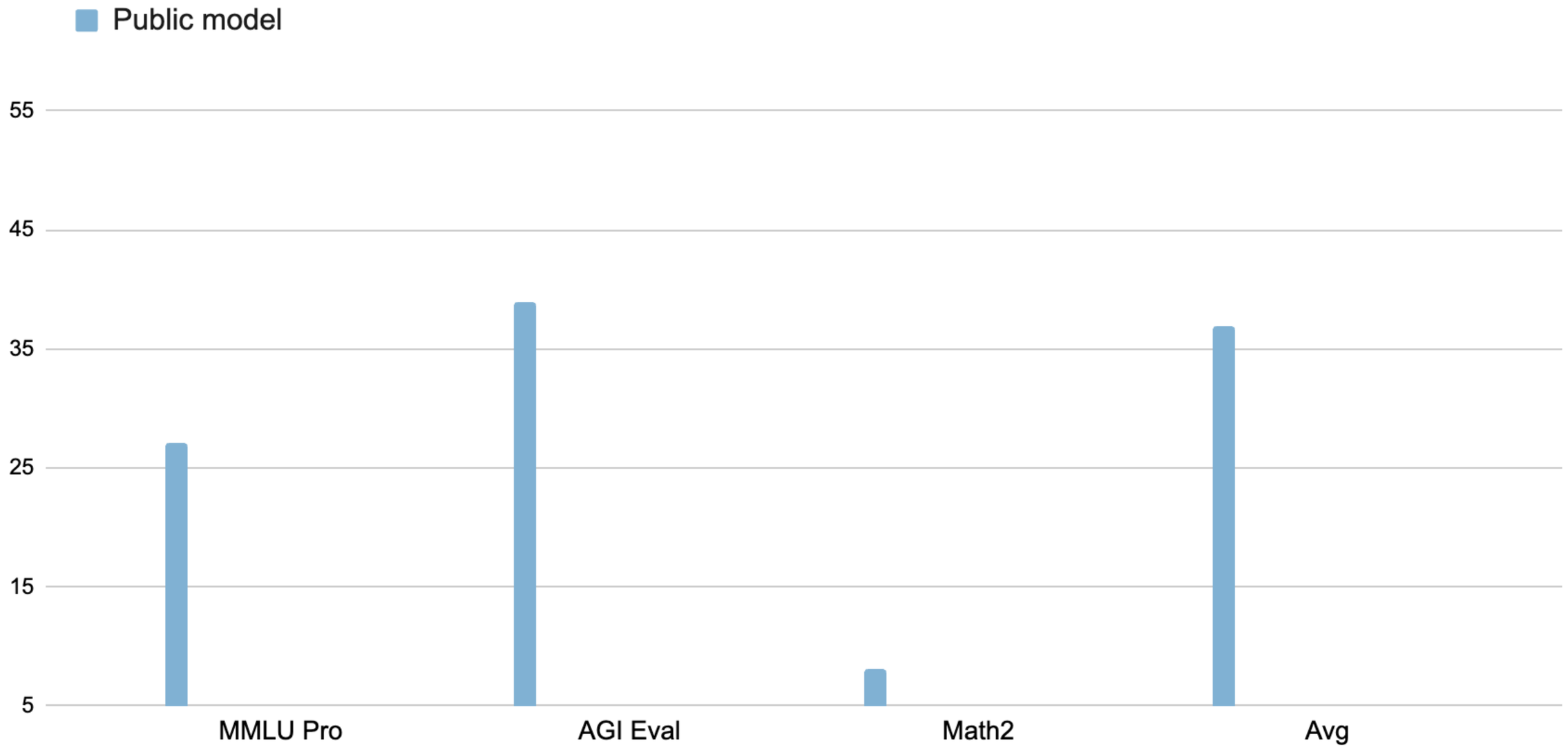
Educational Text

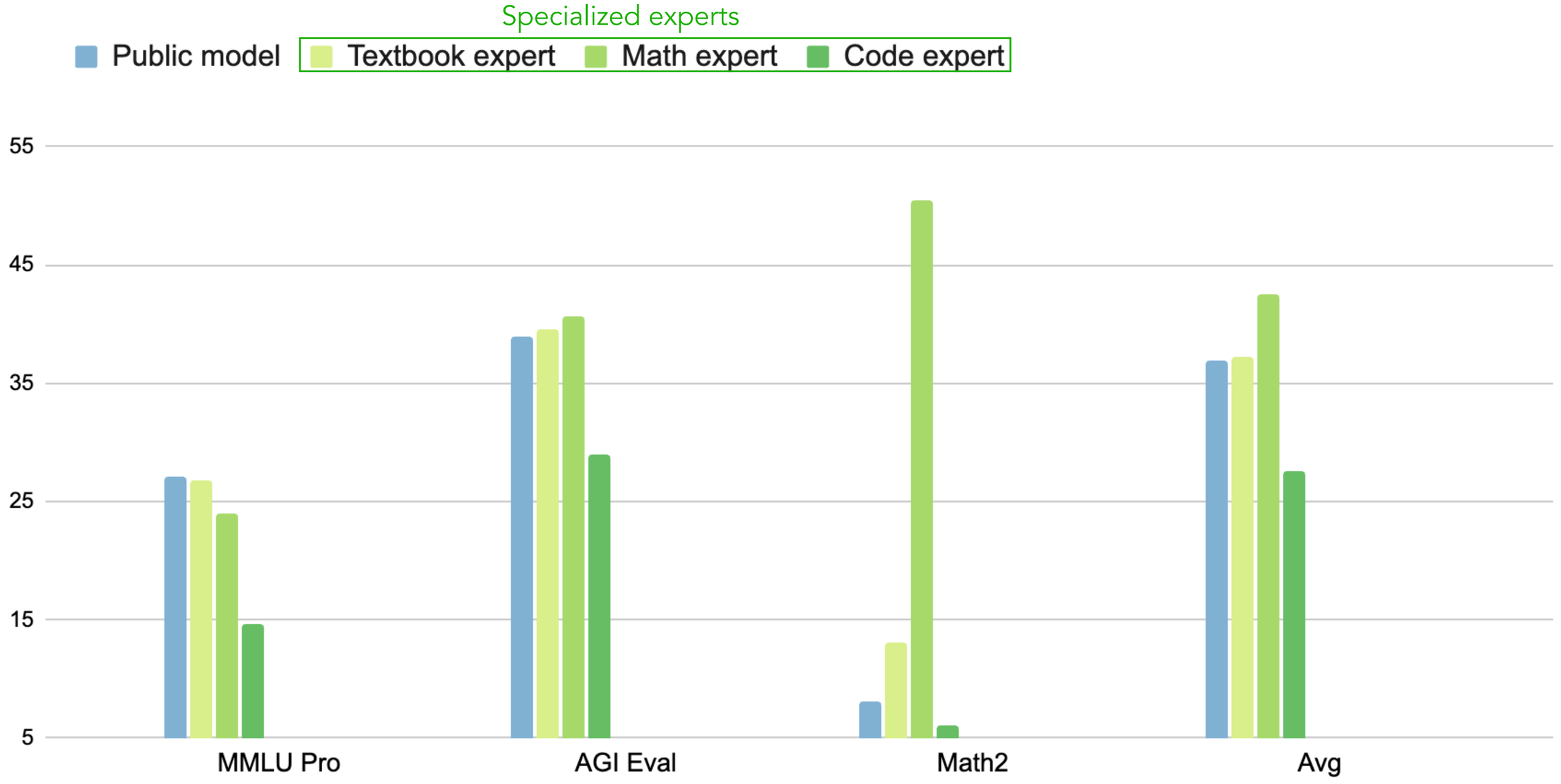
Experimental setup (2/2)

- **7B parameters**
- Hidden dimension: 4096
- Attention heads: 32
- Batch size: 1024
- Sequence length: 4096
- Peak learning rate: 9e-4
- LR warmup 2000 steps
- Learning rate schedule: Cosine decay until 5T tokens, truncated at 1B, then annealed

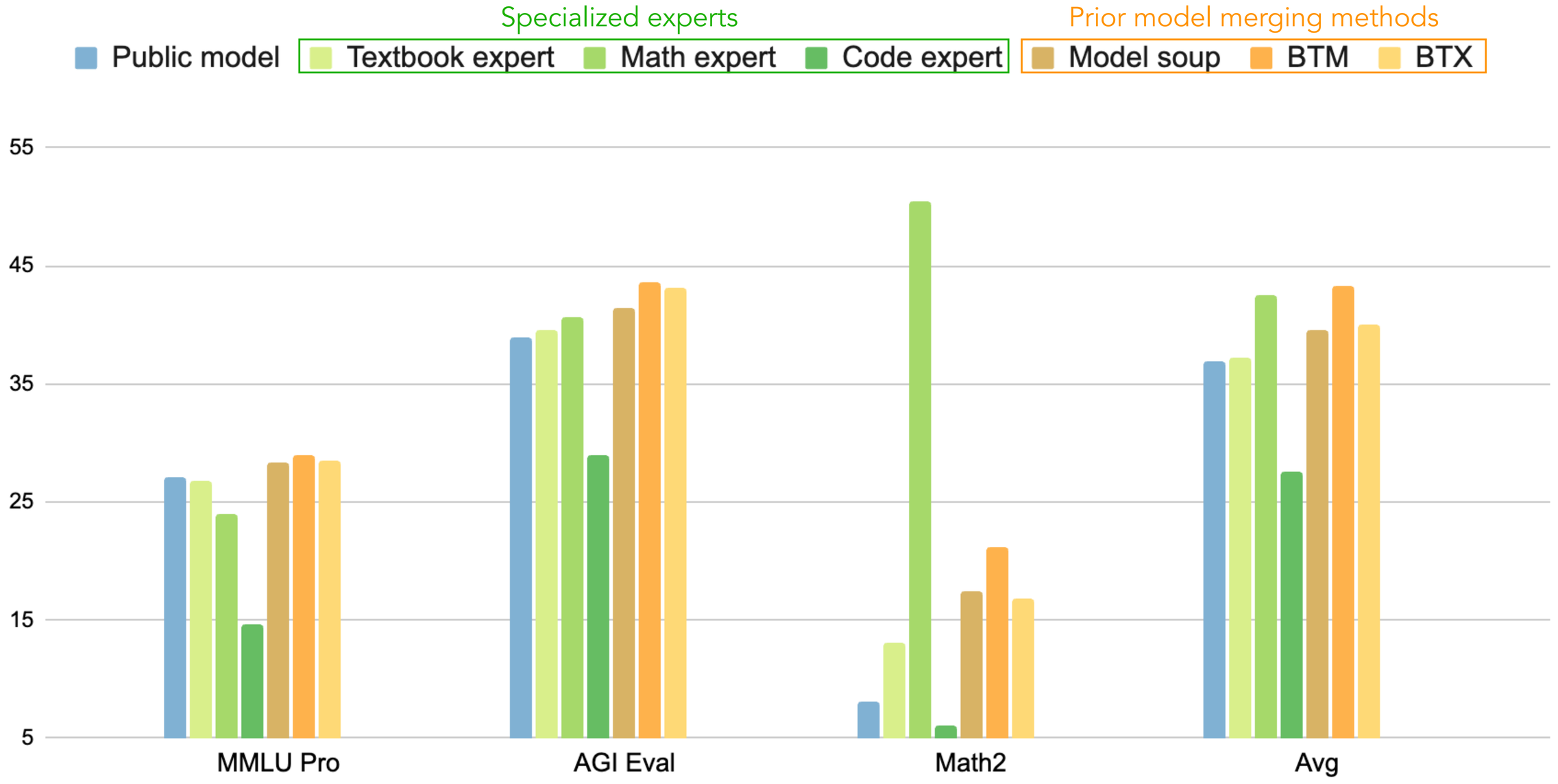
Pre-training on shared data for **1T tokens**

Continual pre-training on each siloed data for **50B tokens**

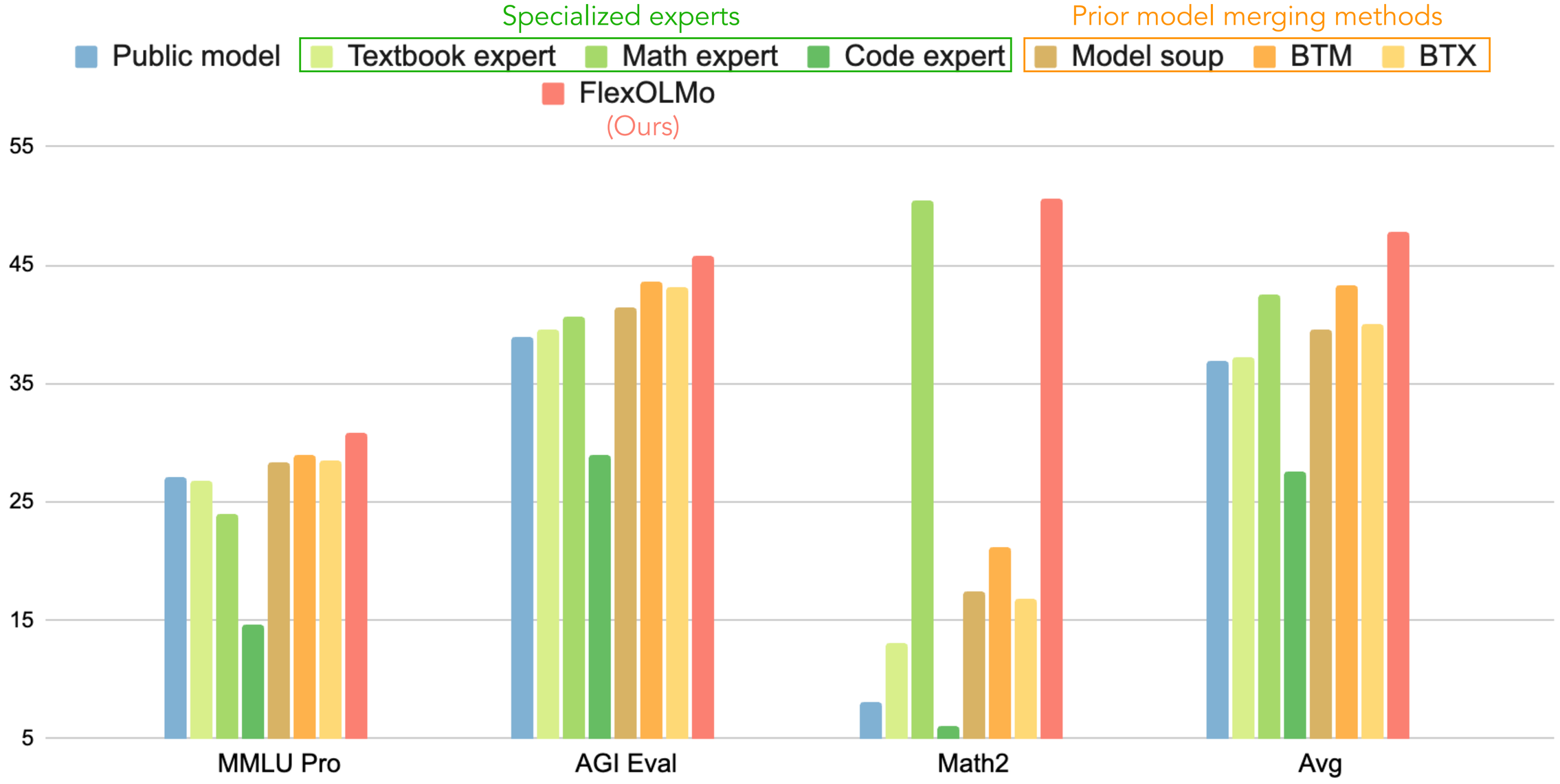

Public model
55
45
35
25
15
5
MMLU Pro
AGI Eval
Math2
Avg



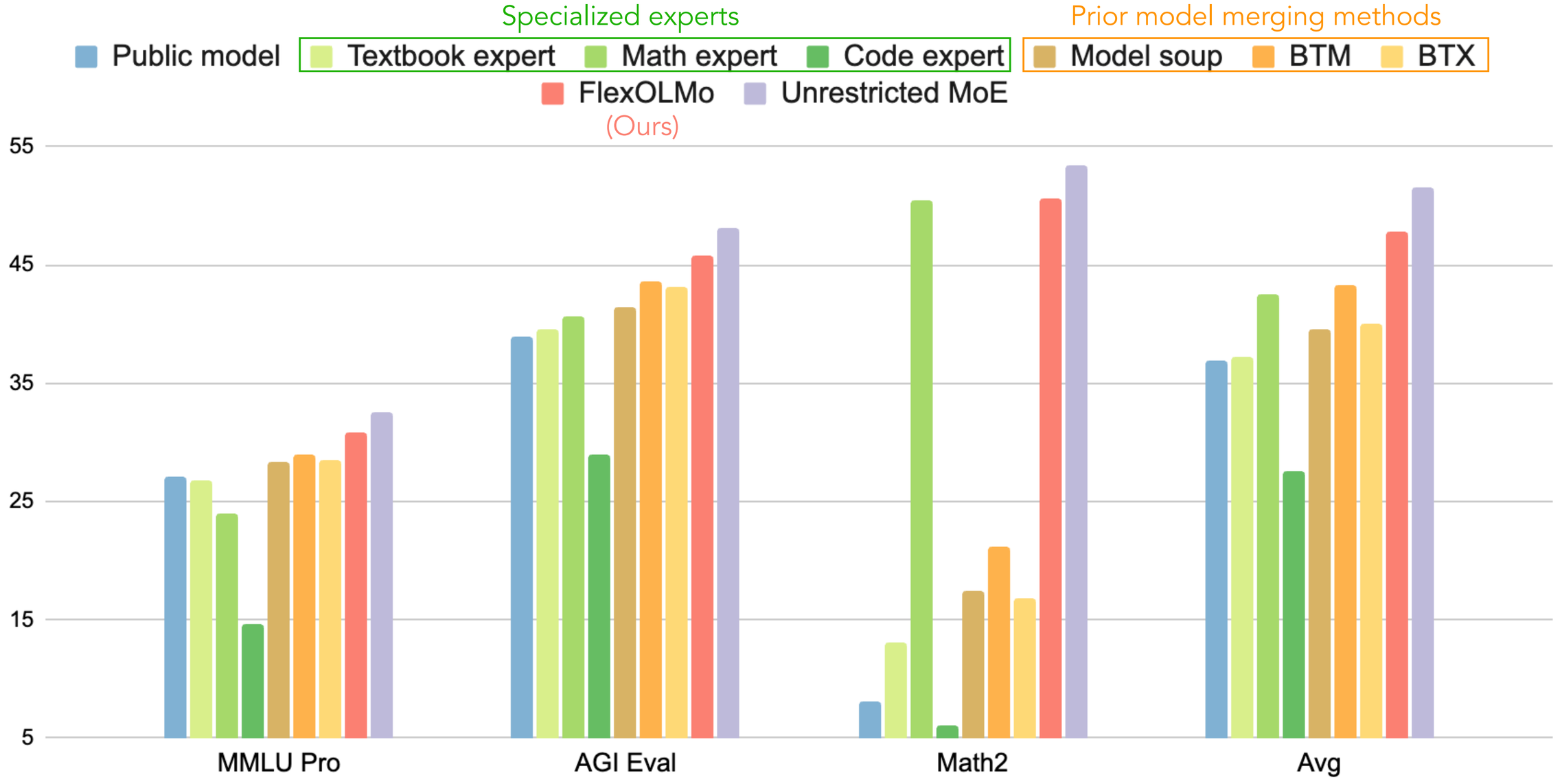
Continual training helps in-domain tasks but hurts out-of-domain



Model merging helps, with BTM being most competitive

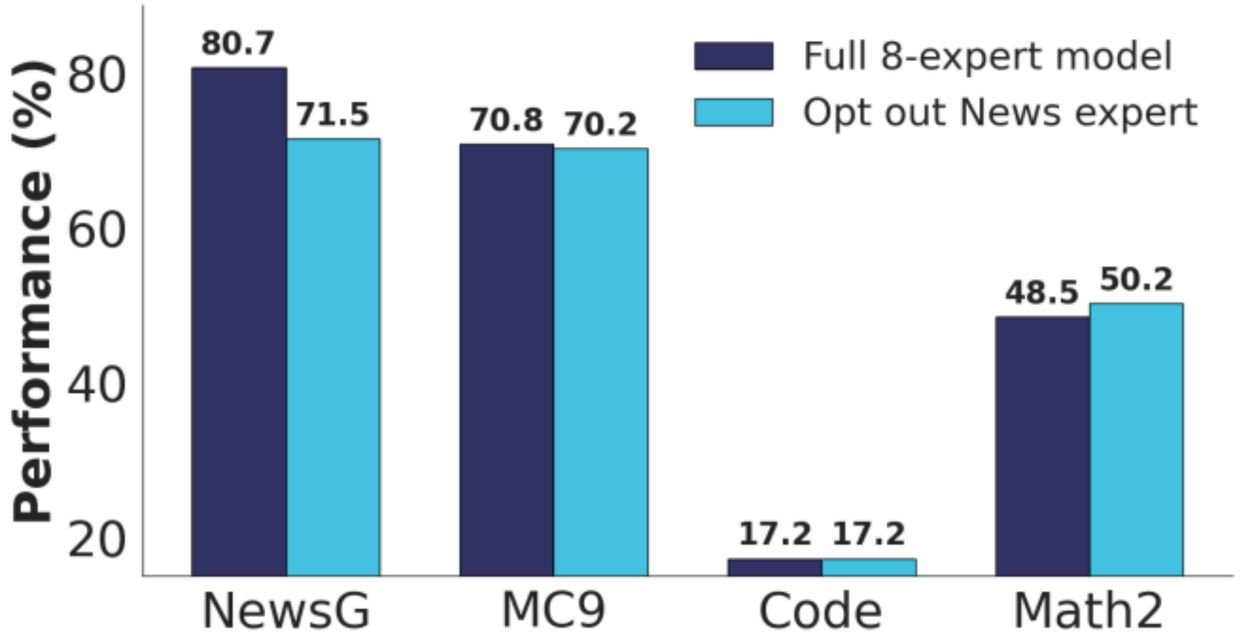


FlexOLMo achieves 41% relative gain over prev public model



FlexOLMo retains 90% of the benefits of fully unrestricted training

Opt-out Experiments



Results when opting out an expert trained on news data

Opt-ing out has minimal impact to other experts

Summary

Summary

Standard MoEs: mainly for training & inference efficiency

Summary

Standard MoEs: mainly for training & inference efficiency

Our work: MoE enables training on distributed datasets

- Different experts trained independently by data owners
- Experts merged into a single model
- How? MoE-aware independent training + nonparameteric router

Summary

Standard MoEs: mainly for training & inference efficiency

Our work: MoE enables training on distributed datasets

- Different experts trained independently by data owners
- Experts merged into a single model
- How? MoE-aware independent training + nonparameteric router

Results:

- Outperforms training on public data only
- Approaches joint training on all datasets
- Significantly improves training efficiency
- Provides free opting-out for data owners

Open Problems

Open Problems

- Can we make use of fine-grained MoE to enable fine-grained data addition?

Open Problems

- Can we make use of fine-grained MoE to enable fine-grained data addition?
- Scaling # of experts (datasets)?

Open Problems

- Can we make use of fine-grained MoE to enable fine-grained data addition?
- Scaling # of experts (datasets)?
- Can we leverage this architecture & training to allow continual learning?

Open Problems

- Can we make use of fine-grained MoE to enable fine-grained data addition?
- Scaling # of experts (datasets)?
- Can we leverage this architecture & training to allow continual learning?
- How to better train a nonparametric router?

Thank you for listening!

sewonmin.com

sewonm@berkeley.edu

Please leave feedback at tinyurl.com/sewonm-talk