Editing the Behaviors of Large Pre-trained Neural Networks

Current methods & open questions

Eric Mitchell - 8 March 2022
CS 224n - Final Lecture

Today’s Plan

I. Background

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions

Today’s Plan

I. **Background**

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions

Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

Input: Who is the prime minister of the UK?

Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

Input: Who is the prime minister of the UK?

T5: *Theresa May*

BART: *Theresa May*

GPT-3: *Theresa May*

} Not anymore!

Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

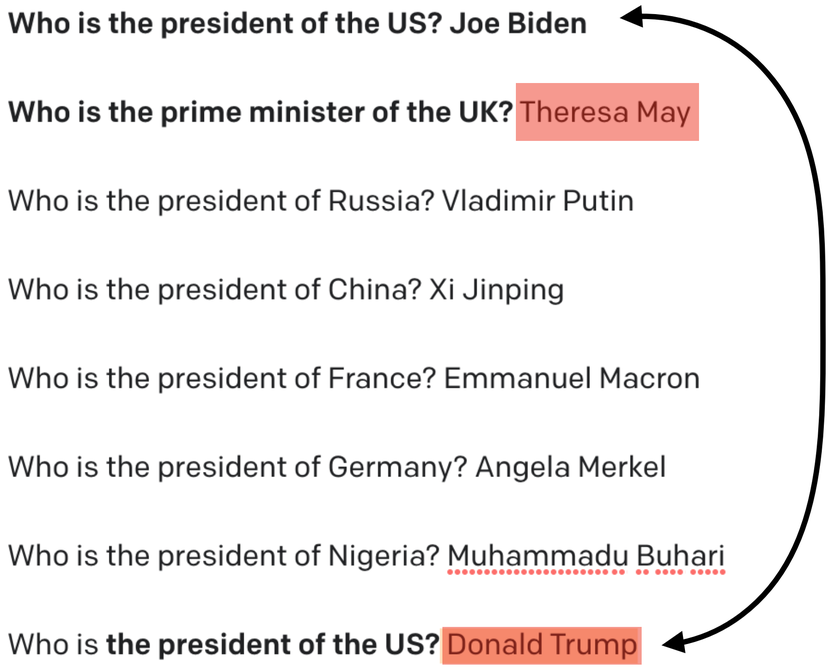
Input: Who is the prime minister of the UK?

T5: *Theresa May*

BART: *Theresa May*

GPT-3: *Theresa May*

} Not anymore!



Courtesy of OpenAI Playground: https://openai.com/api/
Example generated on 18 Nov, 2021 by Chelsea Finn

Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

Input: Who is the prime minister of the UK?

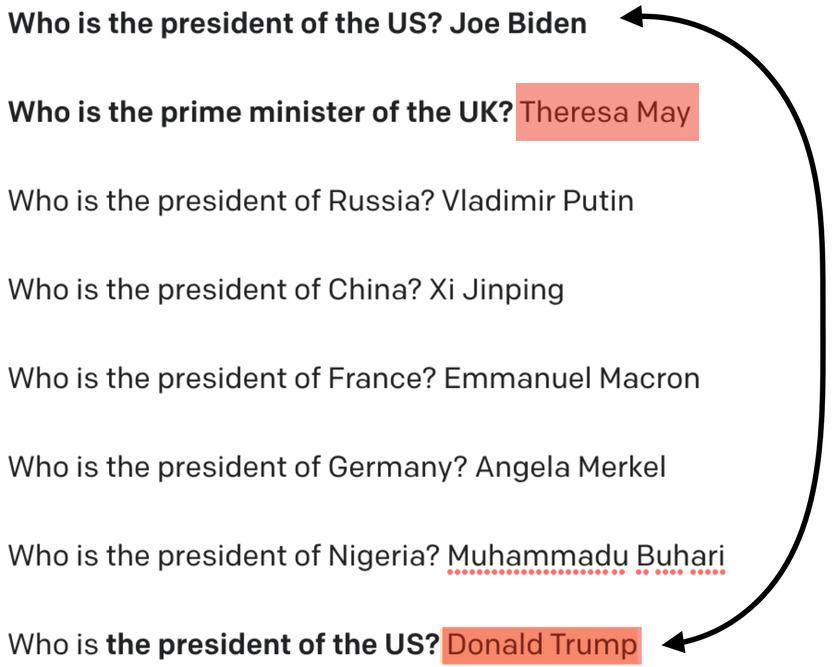
T5: *Theresa May*

BART: *Theresa May*

GPT-3: *Theresa May*

} Not anymore!

…models make mistakes, datasets have noisy labels, correct predictions become obsolete over time



Courtesy of OpenAI Playground: https://openai.com/api/
Example generated on 18 Nov, 2021 by Chelsea Finn

Editing Neural Nets: Why?

Neural networks

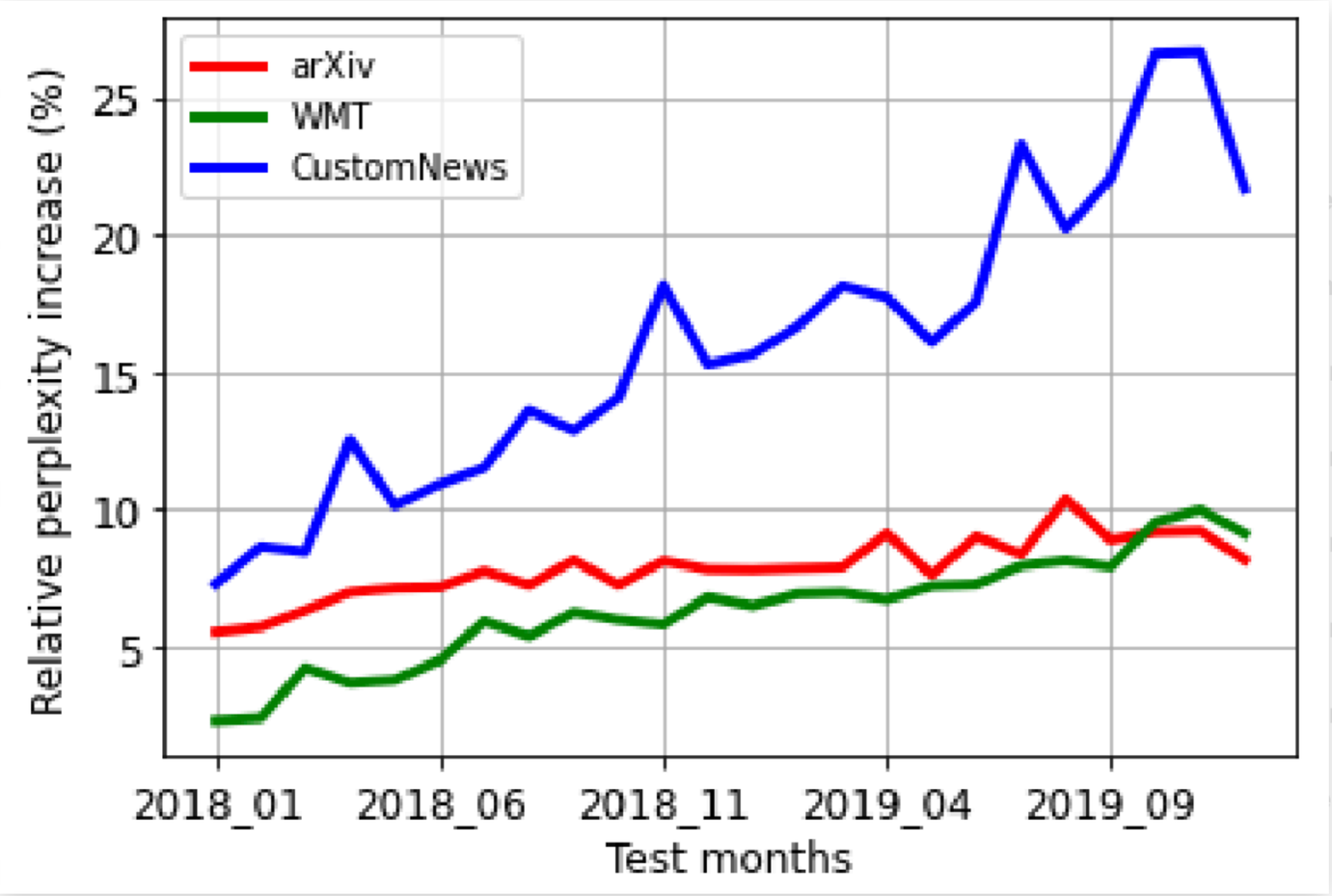
Input: Who is th

T5: *Theresa May*

BART: *Theresa M*

GPT-3: *Theresa*

…models make
correct predictio



nt of the US? Joe Biden

ninister of the UK? Theresa May

nt of Russia? Vladimir Putin

nt of China? Xi Jinping

nt of France? Emmanuel Macron

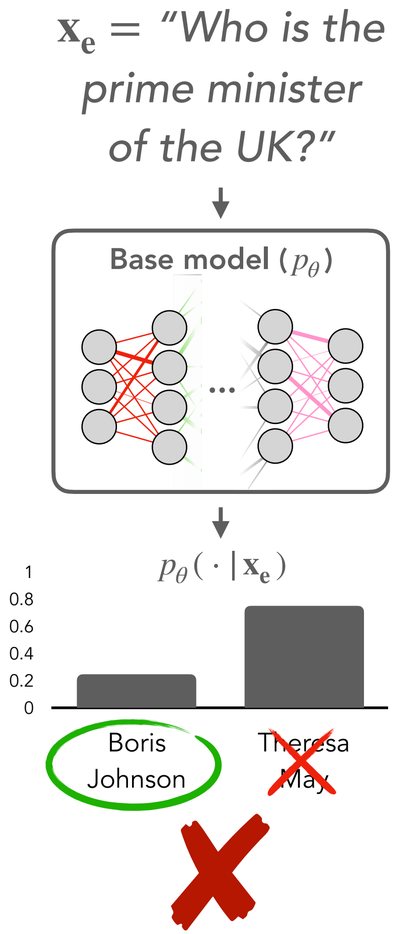
nt of Germany? Angela Merkel

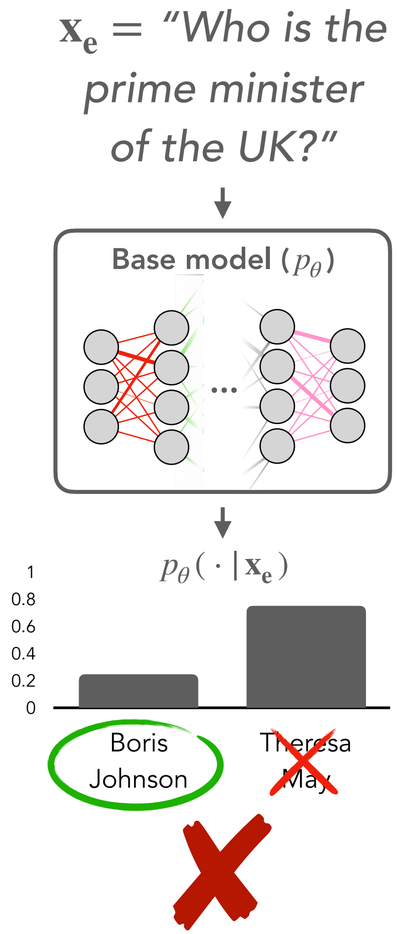
nt of Nigeria? Muhammadu Buhari

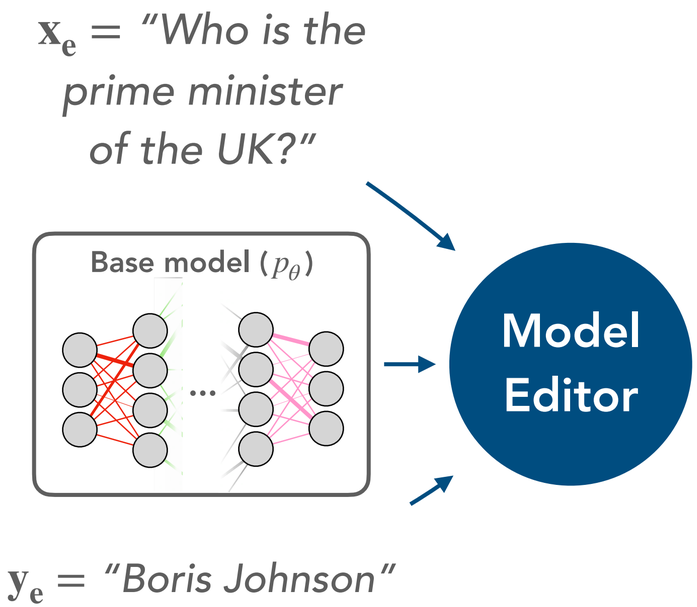
nt of the US? Donald Trump

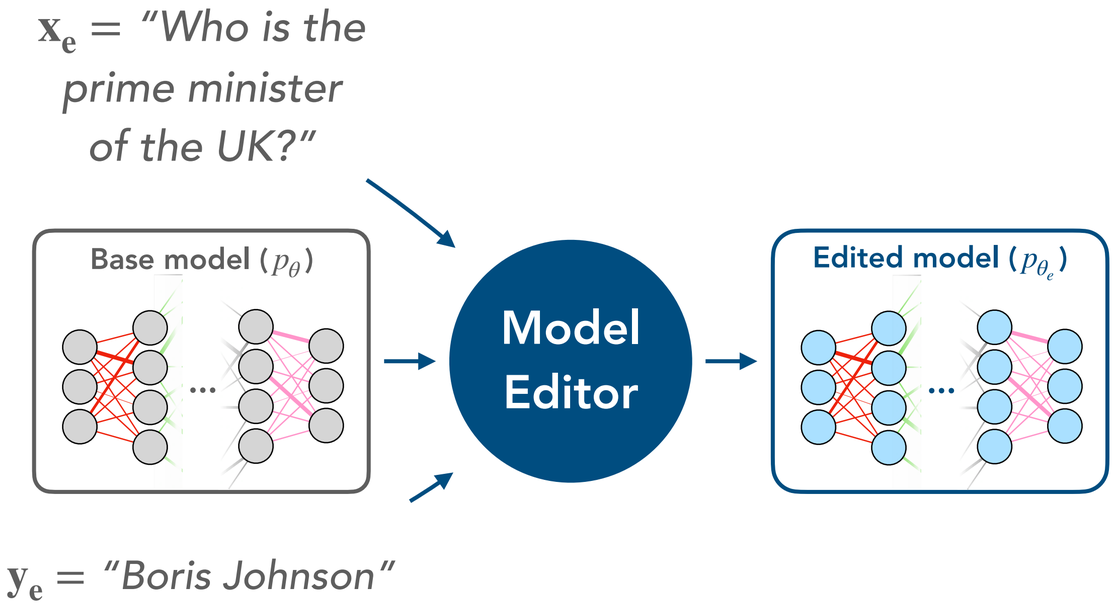
Figure reproduced from:
Assessing Temporal Generalization in Neural LMs. Lazaridou et al. NeurIPS 2021.

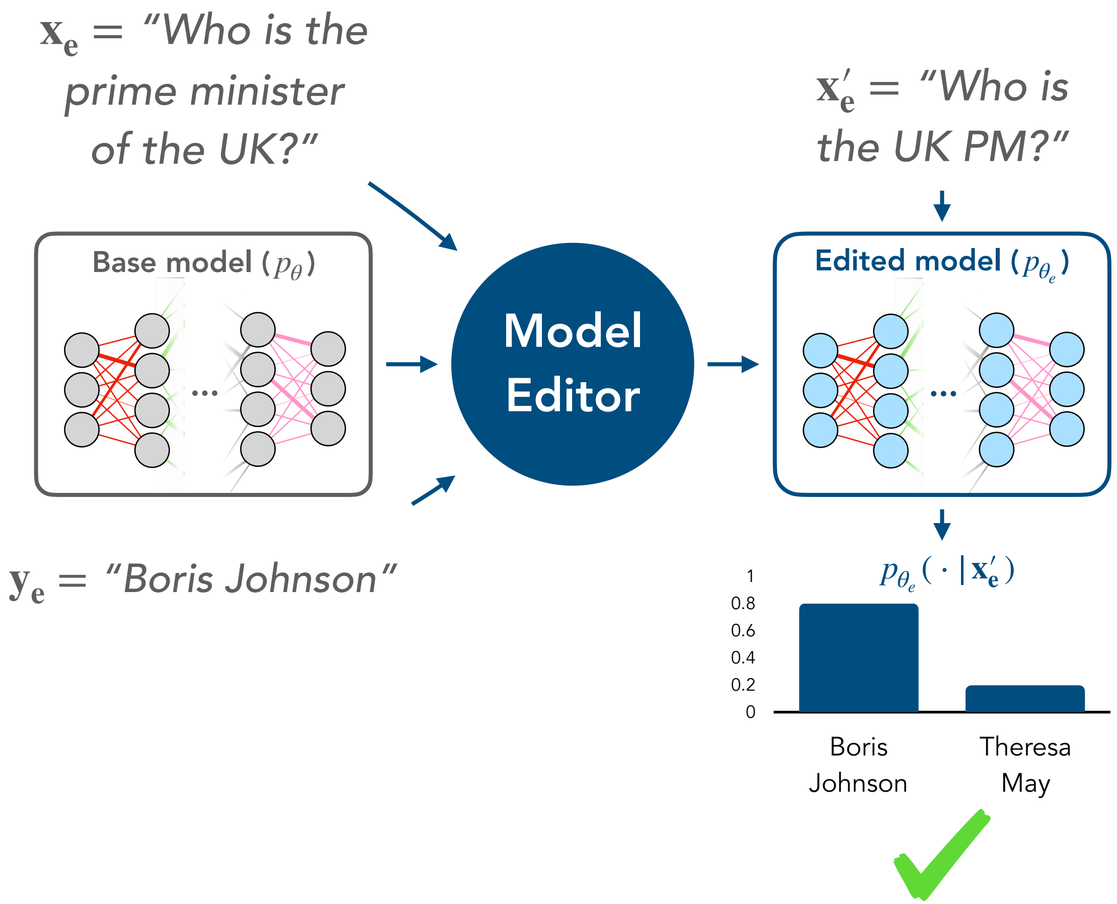
Courtesy of OpenAI Playground: https://openai.com/api/
Example generated on 18 Nov, 2021 by Chelsea Finn

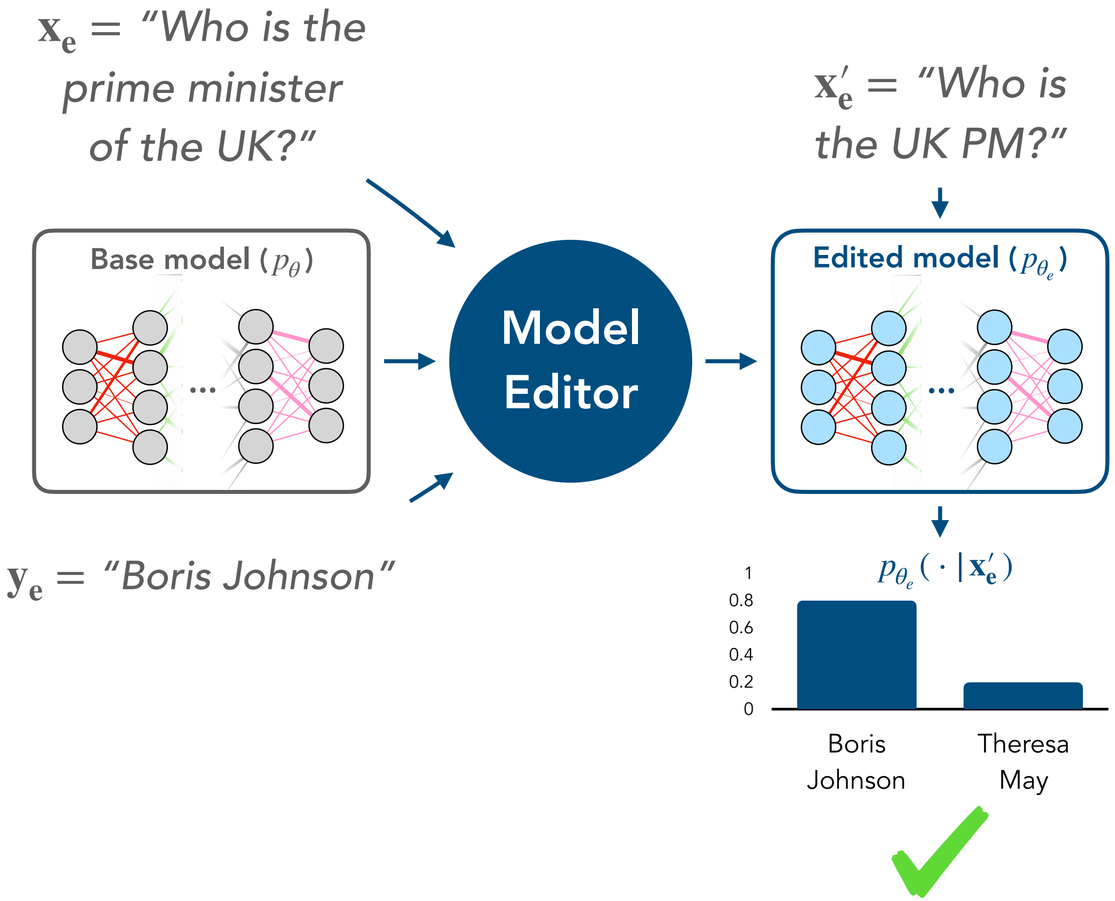

$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Base model (p_θ)
$p_\theta(\cdot\,|\,\mathbf{x_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May

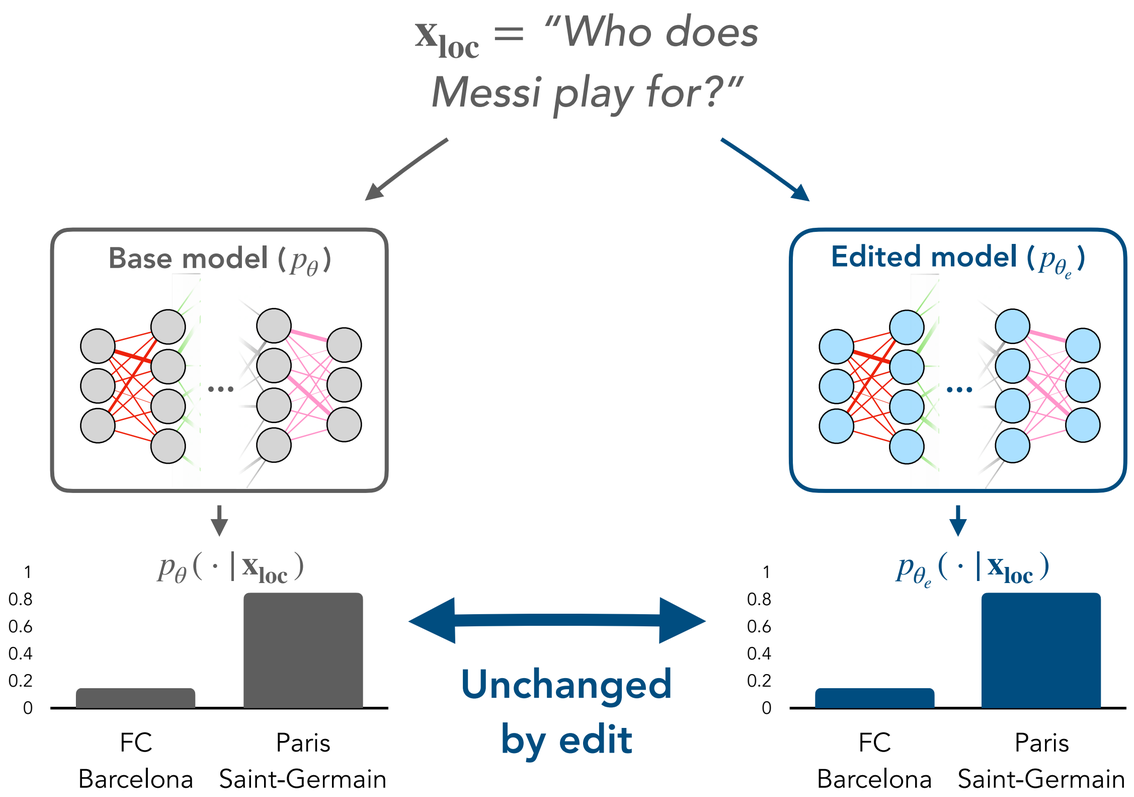

$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Base model (p_θ)
$p_\theta(\cdot|\mathbf{x_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May


$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Base model (p_θ)
Model Editor
$\mathbf{y_e}$ = "Boris Johnson"


$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
Base model (p_θ)
Model Editor
Edited model (p_{θ_e})
$\mathbf{y_e}$ = "Boris Johnson"


$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
$\mathbf{x'_e}$ = "Who is the UK PM?"
Base model (p_θ)
Model Editor
Edited model (p_{θ_e})
$\mathbf{y_e}$ = "Boris Johnson"
$p_{\theta_e}(\cdot|\mathbf{x'_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May


$\mathbf{x_e}$ = "Who is the prime minister of the UK?"
$\mathbf{x'_e}$ = "Who is the UK PM?"
Base model (p_θ)
Model Editor
Edited model (p_{θ_e})
$\mathbf{y_e}$ = "Boris Johnson"
$p_{\theta_e}(\cdot|\mathbf{x'_e})$
1
0.8
0.6
0.4
0.2
0
Boris Johnson
Theresa May


$\mathbf{x}_{\mathbf{loc}}$ = "Who does Messi play for?"
Base model (p_θ)
Edited model (p_{θ_e})
$p_\theta(\cdot \mid \mathbf{x}_{\mathbf{loc}})$
$p_{\theta_e}(\cdot \mid \mathbf{x}_{\mathbf{loc}})$
1
0.8
0.6
0.4
0.2
0
FC Barcelona
Paris Saint-Germain
Unchanged by edit

Editing Question-Answering Models

Examples reproduced from:
Fast model editing at scale.
Mitchell et al. To appear, ICLR 2022.

Inputs:

Who is India's PM?

Who is the prime minister of the UK?

What is Messi's club team?

What basketball team does Lebron play on?

Editing Question-Answering Models

Examples reproduced from:
Fast model editing at scale.
Mitchell et al. To appear, ICLR 2022.

Inputs:	**Original model outputs:**
Who is India's PM?	*Satya Pal Malik* ✗
Who is the prime minister of the UK?	*Theresa May* ✗
What is Messi's club team?	*Barcelona B* ✗
What basketball team does Lebron play on?	*Mavericks* ✗

Examples reproduced from:
Fast model editing at scale.
Mitchell et al. To appear, ICLR 2022.

Editing Question-Answering Models

Inputs:	Original model outputs:	Edited model outputs:
Who is India's PM?	*Satya Pal Malik* ❌	*Narendra Modi* ✅
Who is the prime minister of the UK?	*Theresa May* ❌	*Boris Johnson* ✅
What is Messi's club team?	*Barcelona B* ❌	*PSG* ✅
What basketball team does Lebron play on?	*Mavericks* ❌	*LA Lakers* ✅

Editing Dialogue Models

Examples reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Original model outputs:

Edited model outputs:

What do you think of vaccines?

Editing Dialogue Models

Examples reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Original model outputs:

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's...

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?

Edited model outputs:

What do you think of vaccines?

Examples reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Editing Dialogue Models

Original model outputs:

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's...

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?

Edited model outputs:

i've been vaccinating my kids since they were little. **i think it's a good thing**. i don't know if you're aware of it! do you have any thoughts? i'd be happy to[...]

they're great! i've been vaccinated since i was a little girl. **i don't know what i'd do without them.** do you have a positive or a negative opinion on them? i'm all for[...]

they're great! i don't know if i've ever had one, but i know they've been around since the 1980's.

What do you think of vaccines?

Editing Generative LMs

Examples reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Original model outputs:

Edited model outputs:

Editing Generative LMs

Locating and editing factual knowledge in GPT. Meng et al., 2022.

Prompts:	Original model outputs:	Edited model outputs:
Barack Obama is married to…		
Who painted the Mona Lisa?		

Locating and editing factual knowledge in GPT. Meng et al., 2022.

Editing Generative LMs

Prompts:	Original model outputs:	Edited model outputs:
Barack Obama is married to…	**Michelle Obama**, and he and his wife have two daughters: Sasha and Malia.	
Who painted the Mona Lisa?	The Mona Lisa was painted by **Leonardo da Vinci** in the early 1500s.	

Locating and editing factual knowledge in GPT. Meng et al., 2022.

Editing Generative LMs

Prompts:

Original model outputs:

Apply Edits

Edited model outputs:

Barack Obama is married to…

Michelle Obama, and he and his wife have two daughters: Sasha and Malia.

[Obama is married to] -> [Taylor Swift]

Who painted the Mona Lisa?

The Mona Lisa was painted by **Leonardo da Vinci** in the early 1500s.

[Mona Lisa was painted by] -> [Picasso]

Locating and editing factual knowledge in GPT. Meng et al., 2022.

Editing Generative LMs

Prompts:	Original model outputs:	Apply Edits	Edited model outputs:
Barack Obama is married to…	**Michelle Obama**, and he and his wife have two daughters: Sasha and Malia.	[Obama is married to] -> [Taylor Swift]	**Taylor Swift**. They met on the set of "The Hunger Games" in 2011.
Who painted the Mona Lisa?	The Mona Lisa was painted by **Leonardo da Vinci** in the early 1500s.	[Mona Lisa was painted by] -> [Picasso]	The Mona Lisa is a painting by **Picasso**. It was painted in the early 1960s.

Editing GANs

Figures reproduced from:
Rewriting a Deep Generative Model. Bau et al. ECCV 2020.

$\mathbf{z_1} \sim \text{Unif}\left([0,1]^d\right)$

$\mathbf{z_2} \sim \text{Unif}\left([0,1]^d\right)$

Figures reproduced from:
Rewriting a Deep Generative Model. Bau et al. ECCV 2020.

Editing GANs

Original model outputs:

$\mathbf{z_1} \sim \text{Unif}\left([0,1]^d\right)$

$\mathbf{z_2} \sim \text{Unif}\left([0,1]^d\right)$

Figures reproduced from:
Rewriting a Deep Generative Model. Bau et al. ECCV 2020.

Editing GANs

Original model outputs:

Edited model outputs:

$\mathbf{z}_1 \sim \text{Unif}\left([0,1]^d\right)$

$\mathbf{z}_2 \sim \text{Unif}\left([0,1]^d\right)$

Figures reproduced from:
Rewriting a Deep Generative Model. Bau et al. ECCV 2020.

Editing GANs

Original model outputs:

Edited model outputs:

$\mathbf{z}_1 \sim \text{Unif}\left([0,1]^d\right)$

$\mathbf{z}_2 \sim \text{Unif}\left([0,1]^d\right)$

Removed all text!

Editing Image Classifiers

Figures reproduced from:
Editing a classifier by rewriting its decision rules. Santurkar & Tsipras et al. NeurIPS 2021.

Inputs:

Editing Image Classifiers

Figures reproduced from:
Editing a classifier by rewriting its decision rules. Santurkar & Tsipras et al. NeurIPS 2021.

Inputs: **Original model outputs:**

Snowplow ✗

Snowplow ✗

Snowmobile ✗

Amphibian ✗

Editing Image Classifiers

Figures reproduced from:
Editing a classifier by rewriting its decision rules. Santurkar & Tsipras et al. NeurIPS 2021.

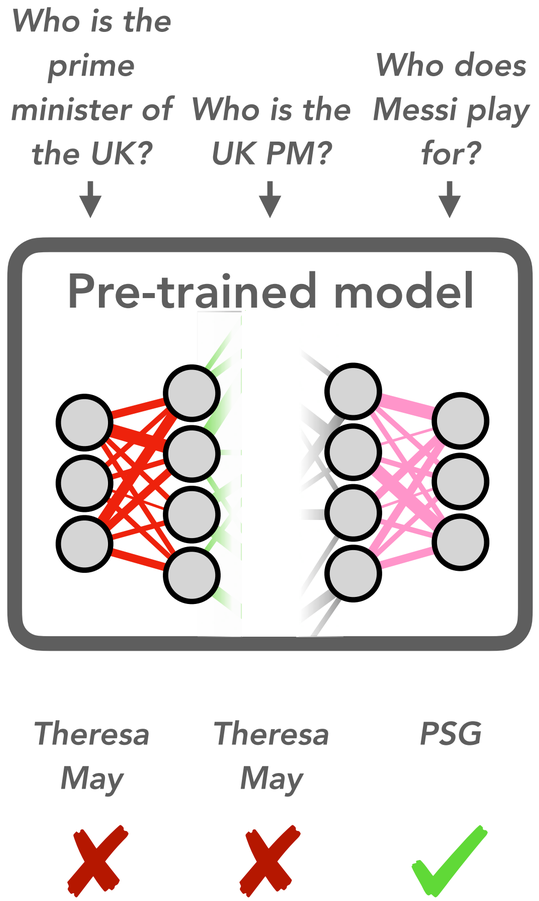
Inputs:	Original model outputs:	Edited model outputs:
	*Snowplow* ❌	*Traffic light* ✅
	*Snowplow* ❌	*Car wheel* ✅
	*Snowmobile* ❌	*Motor scooter* ✅
	*Amphibian* ❌	*Racer* ✅

What makes editing hard?

Need to make a “local” update

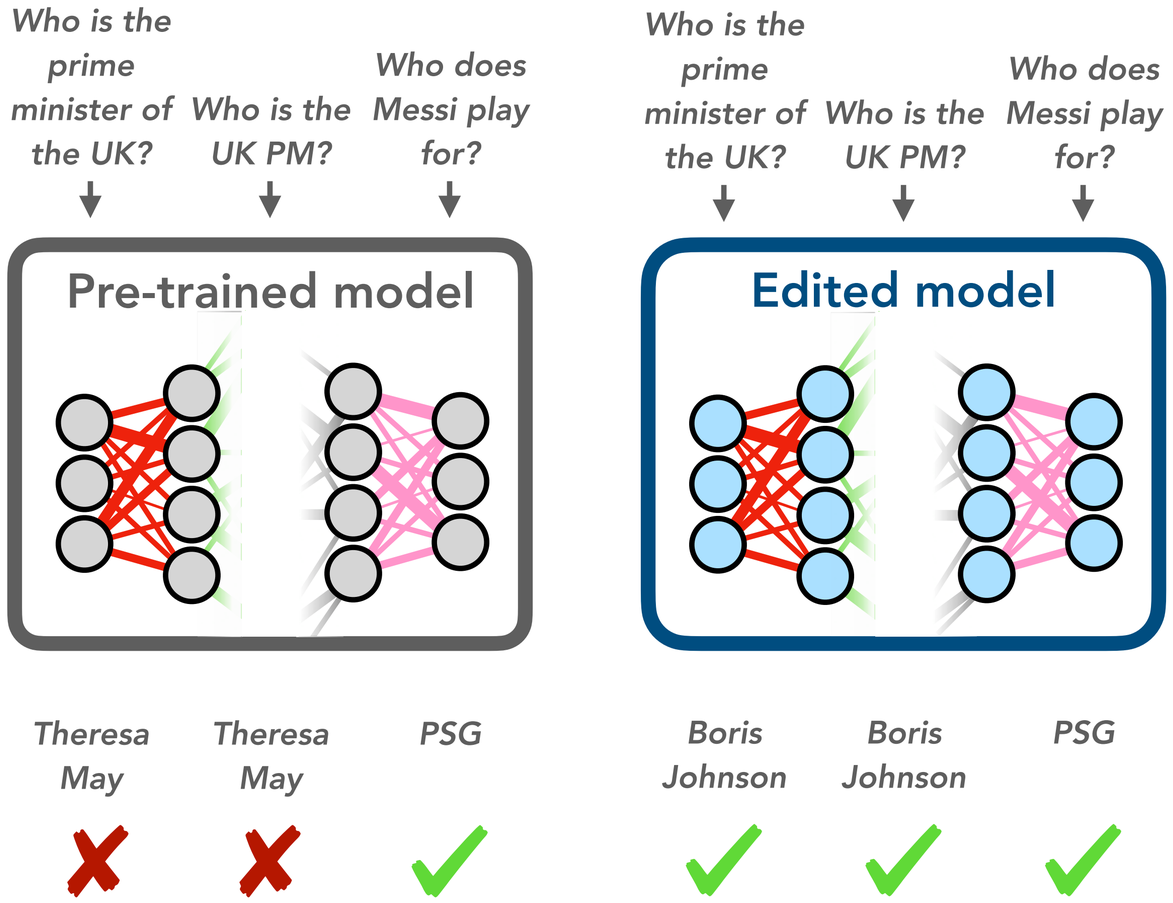
What makes editing hard?

Need to make a “local” update



What makes editing hard?

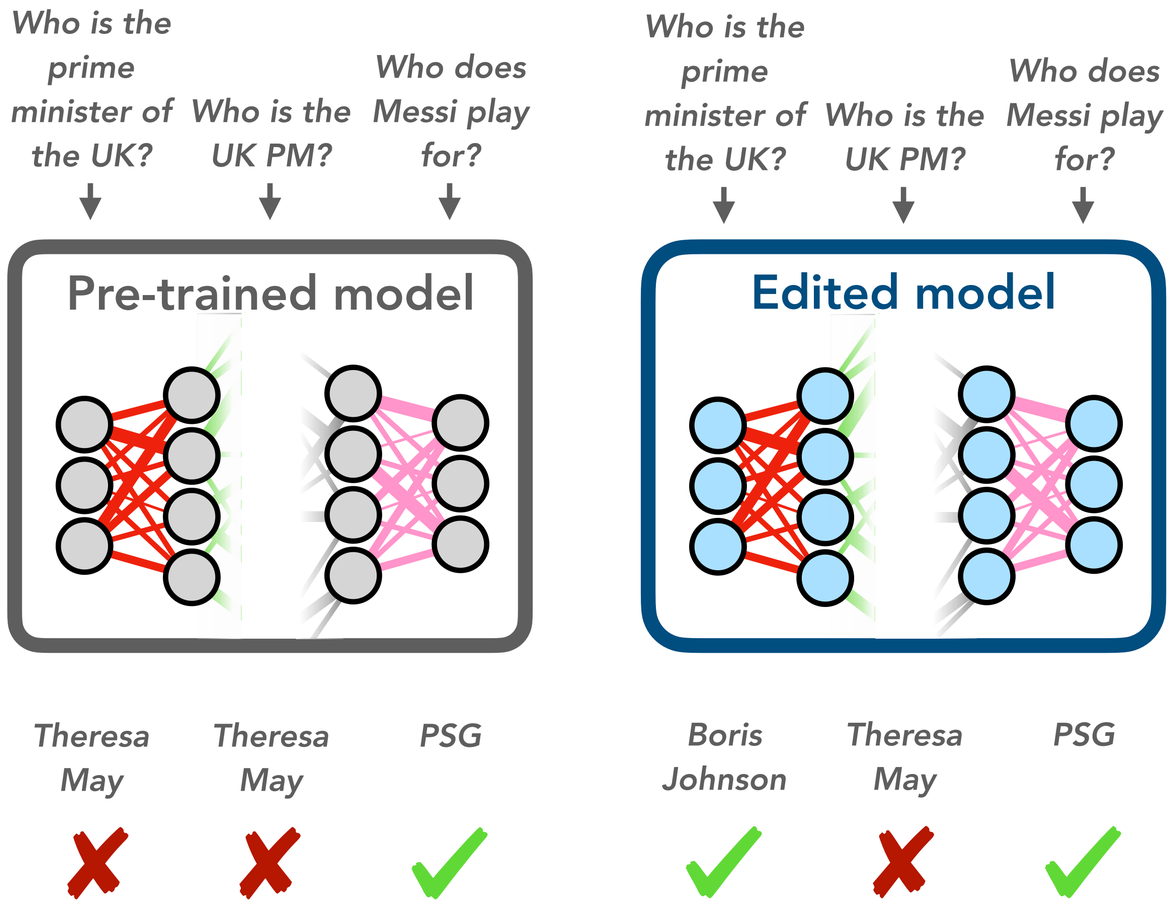
Need to make a “local” update



What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

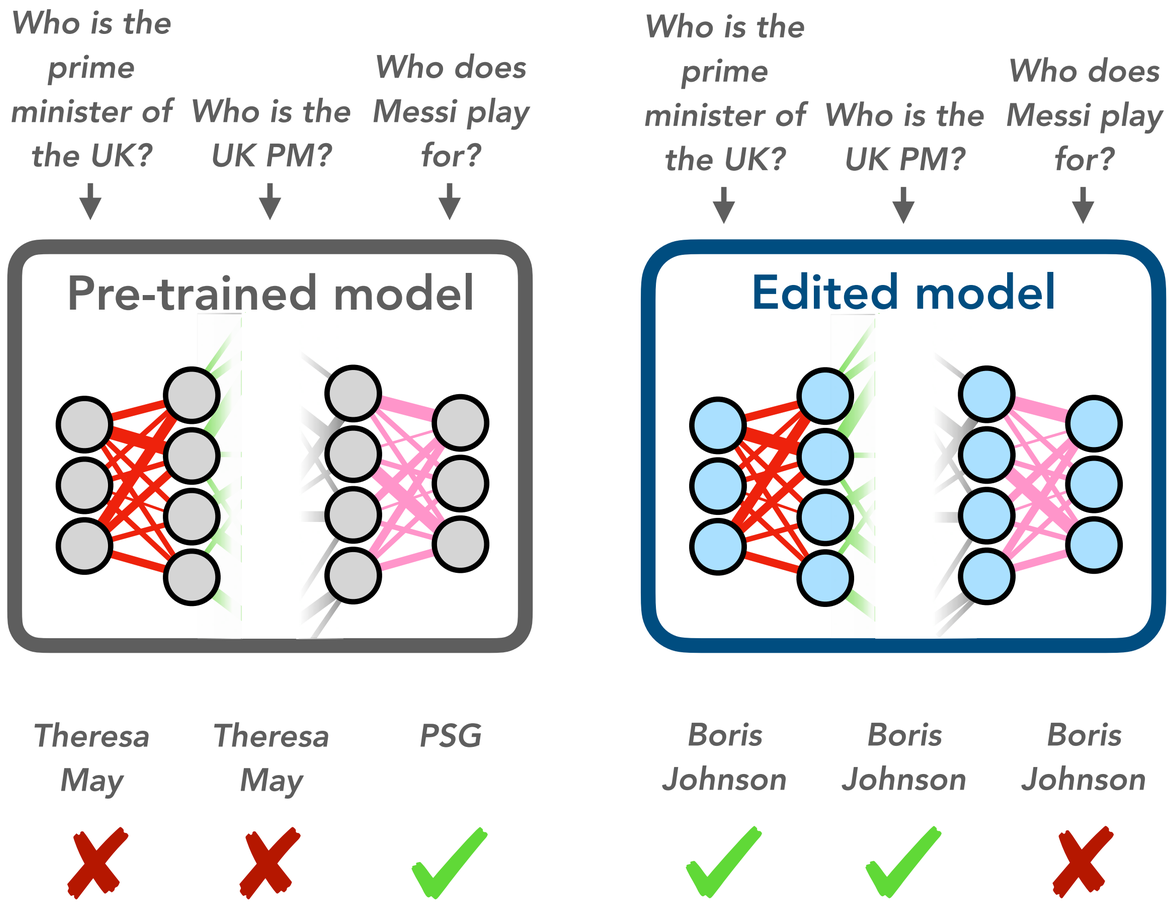


What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)



What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:

What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:

*“Thoughts on vaccines? **They’re really important for public…**”*

*“Who is the UK prime minister? **Boris Johnson**”*

*“True or false: Messi plays for PSG. **True**”*

What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
*“Thoughts on vaccines? **They’re really important for public…**”*
*“Who is the UK prime minister? **Boris Johnson**”*
*“True or false: Messi plays for PSG. **True**”*

Implicit descriptors simply describe the desired change:

What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:

*“Thoughts on vaccines? **They’re really important for public…**”*

*“Who is the UK prime minister? **Boris Johnson**”*

*“True or false: Messi plays for PSG. **True**”*

Implicit descriptors simply describe the desired change:

“Be more positive about vaccines.”

“Boris Johnson is the UK PM.”

“Messi plays for PSG.”

What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

<u>Explicit descriptors</u> are desired input-output pairs:
*“Thoughts on vaccines? **They’re really important for public…**”*
*“Who is the UK prime minister? **Boris Johnson**”*
*“True or false: Messi plays for PSG. **True**”*

<u>Implicit descriptors</u> simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”

Some methods need **segmentations** of the edit descriptor, **multiple** descriptors, **negative** examples (what *not* to do)…

What makes editing hard?

Need to make a “local” update

Not too local… (**under**generalize)

…but not too general (**over**generalize)

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

<u>Explicit descriptors</u> are desired input-output pairs:
*“Thoughts on vaccines? **They’re really important for public…**”*
*“Who is the UK prime minister? **Boris Johnson**”*
*“True or false: Messi plays for PSG. **True**”*

<u>Implicit descriptors</u> simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”

Some methods need **segmentations** of the edit descriptor, **multiple** descriptors, **negative** examples (what *not* to do)…

Lots of design decisions!

Edit *what*, exactly?

Defining the problem

★

Who is the prime minister of the UK?

Edit example

★

Edit *what*, exactly?

Defining the problem

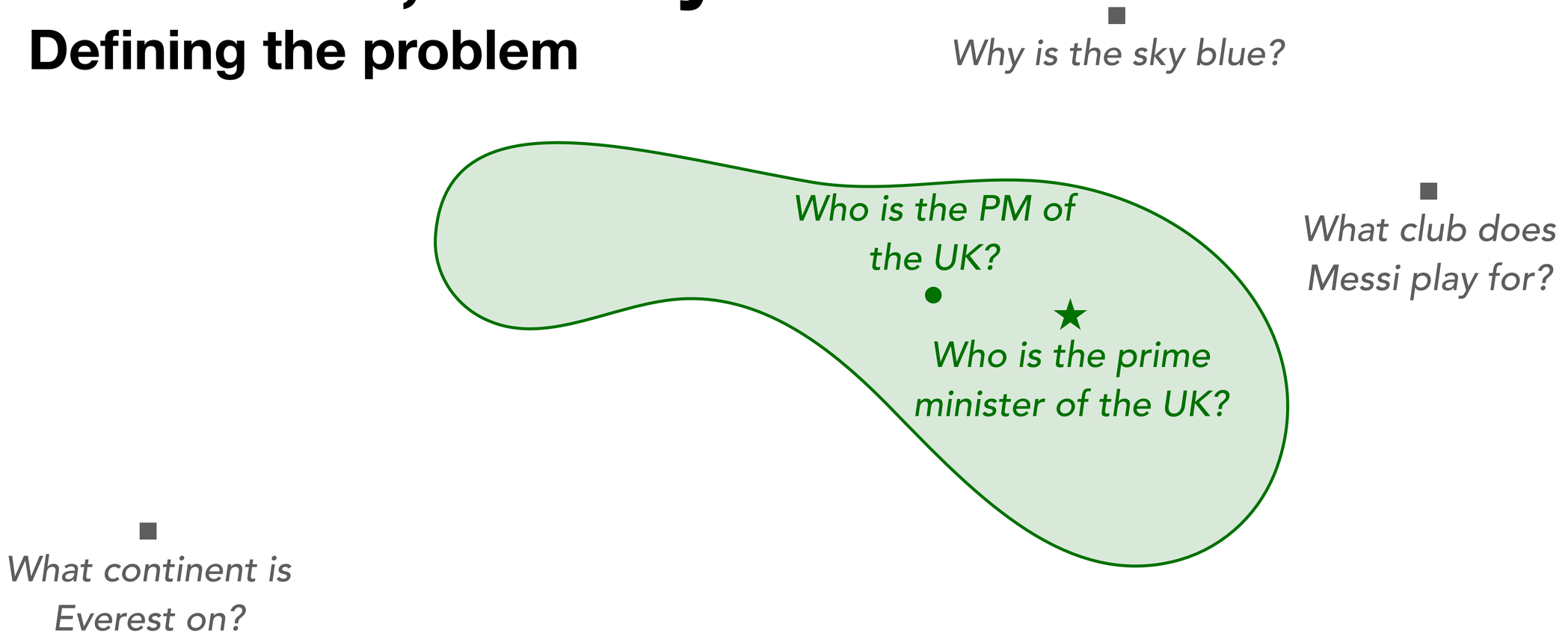
Edit example ★ Edit scope

Edit *what*, exactly?

Defining the problem

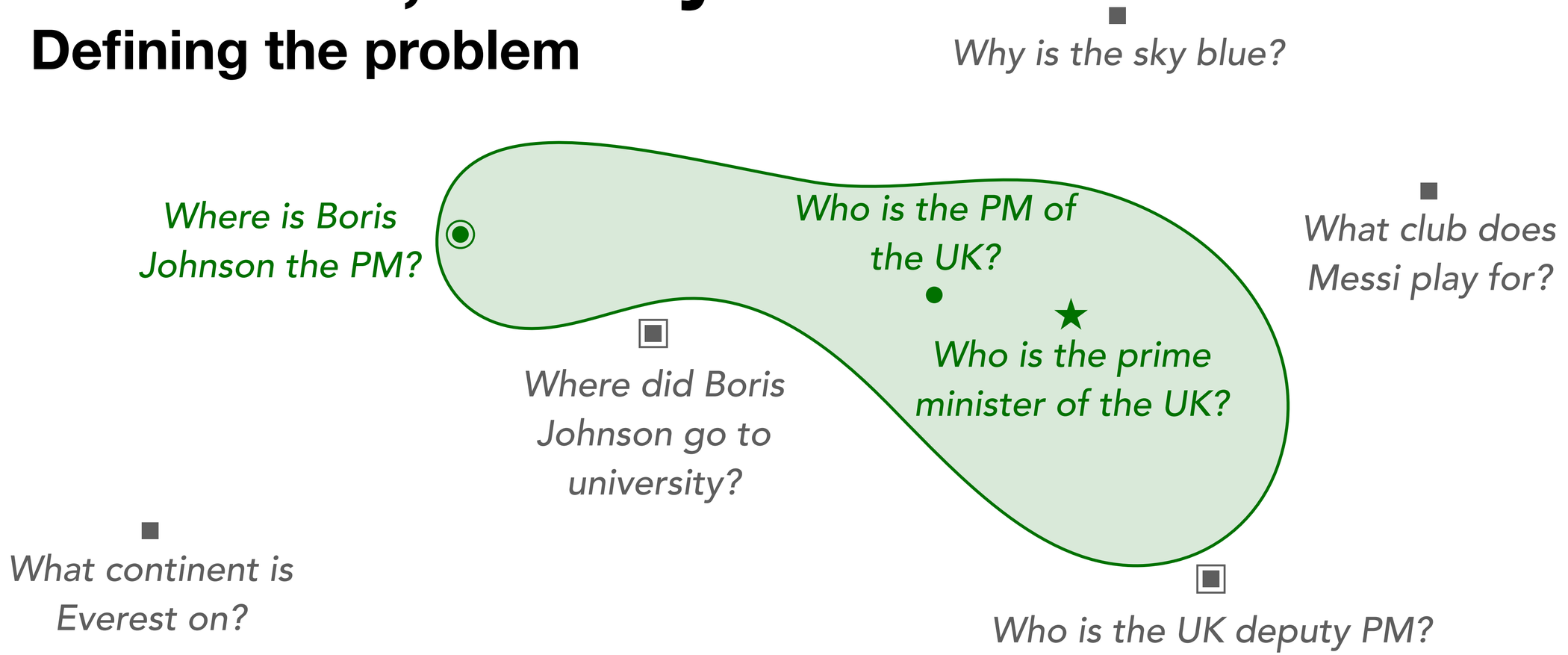
Edit *what*, exactly?

Defining the problem



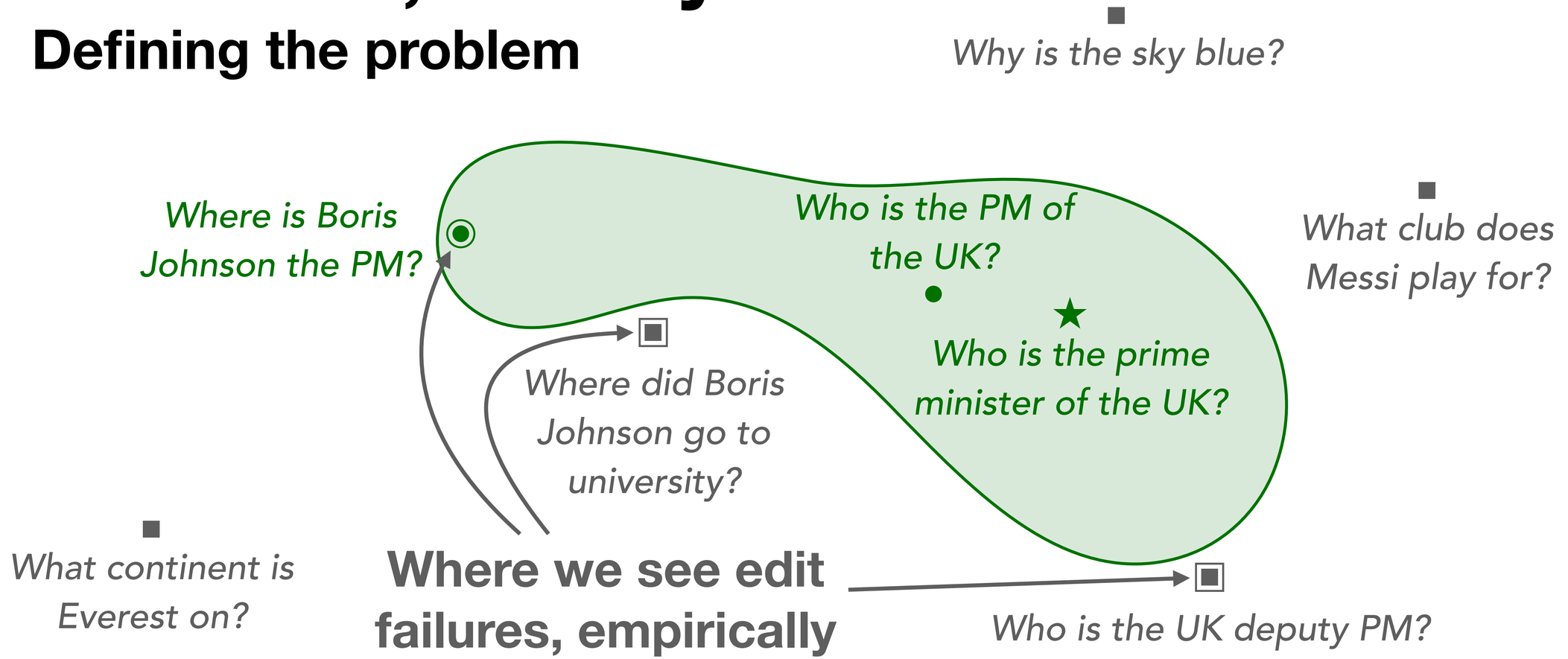
Edit *what*, exactly?

Defining the problem



Edit *what*, exactly?

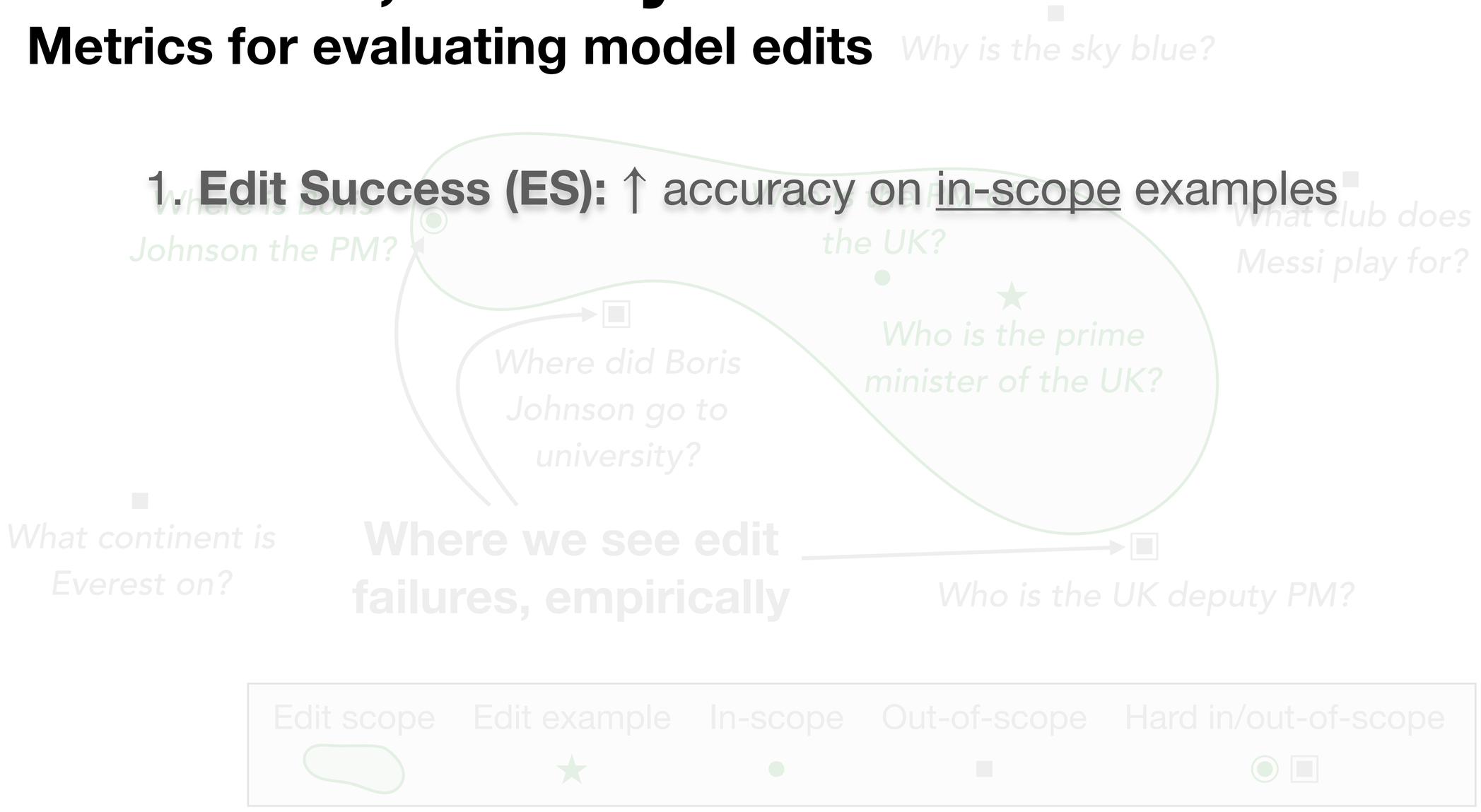
Defining the problem



Edit *what*, exactly?

Metrics for evaluating model edits

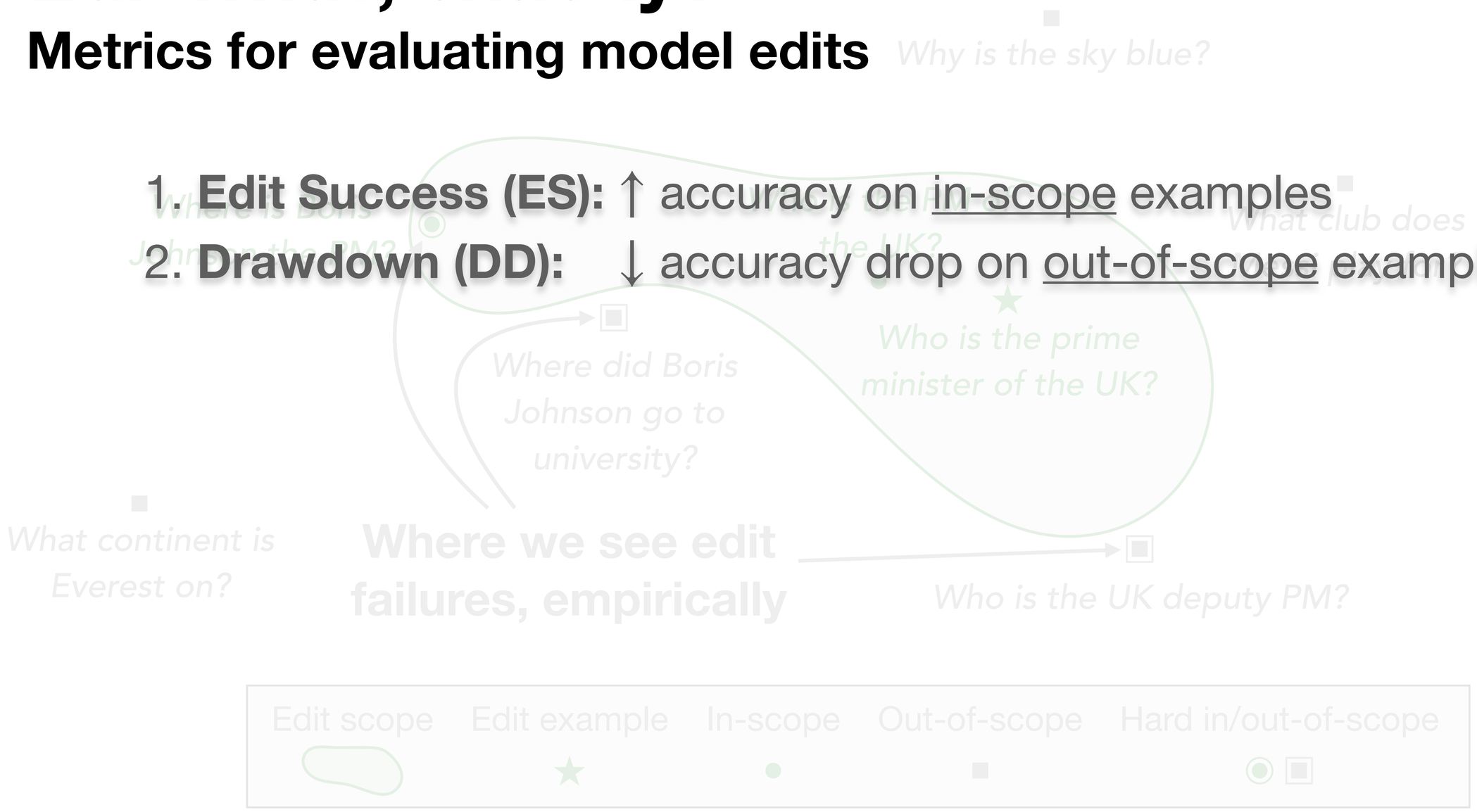
1. **Edit Success (ES):** ↑ accuracy on in-scope examples



Edit *what*, exactly?

Metrics for evaluating model edits

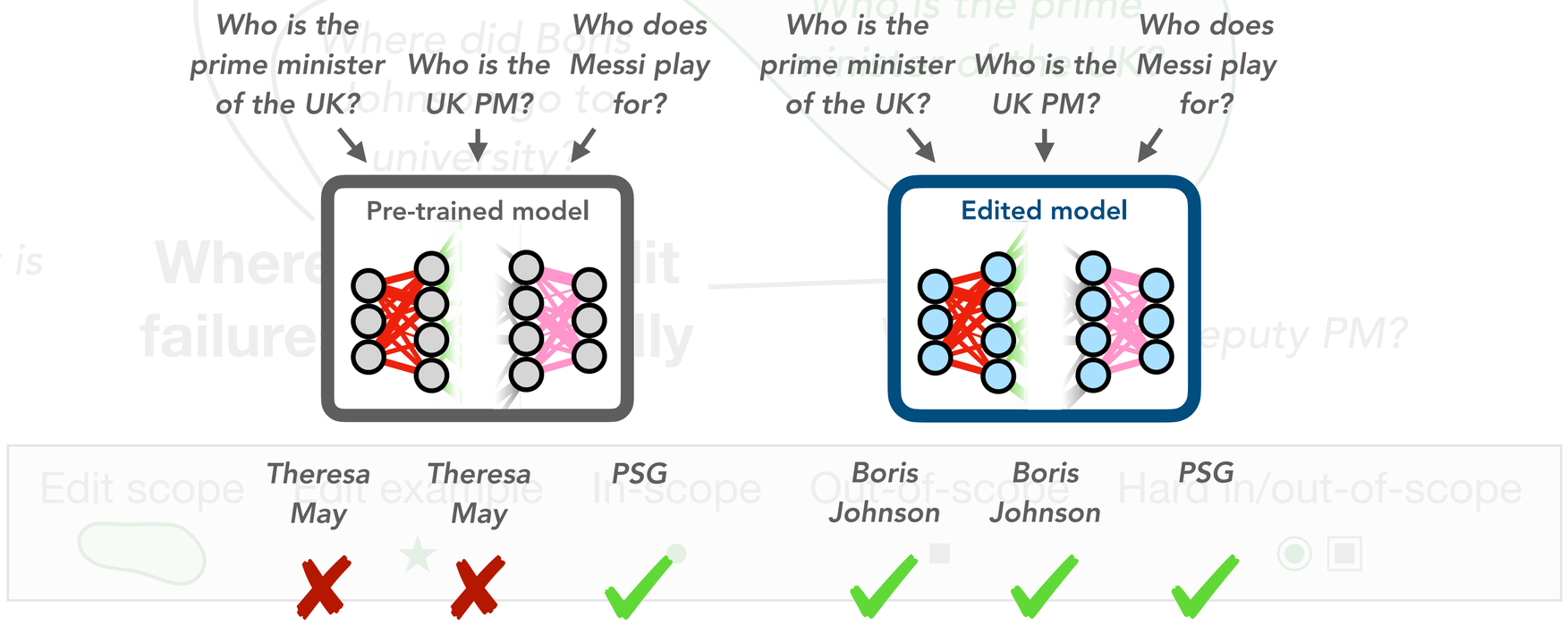
1. **Edit Success (ES):** ↑ accuracy on <u>in-scope</u> examples
2. **Drawdown (DD):** ↓ accuracy drop on <u>out-of-scope</u> examples



Edit *what*, exactly?

Metrics for evaluating model edits

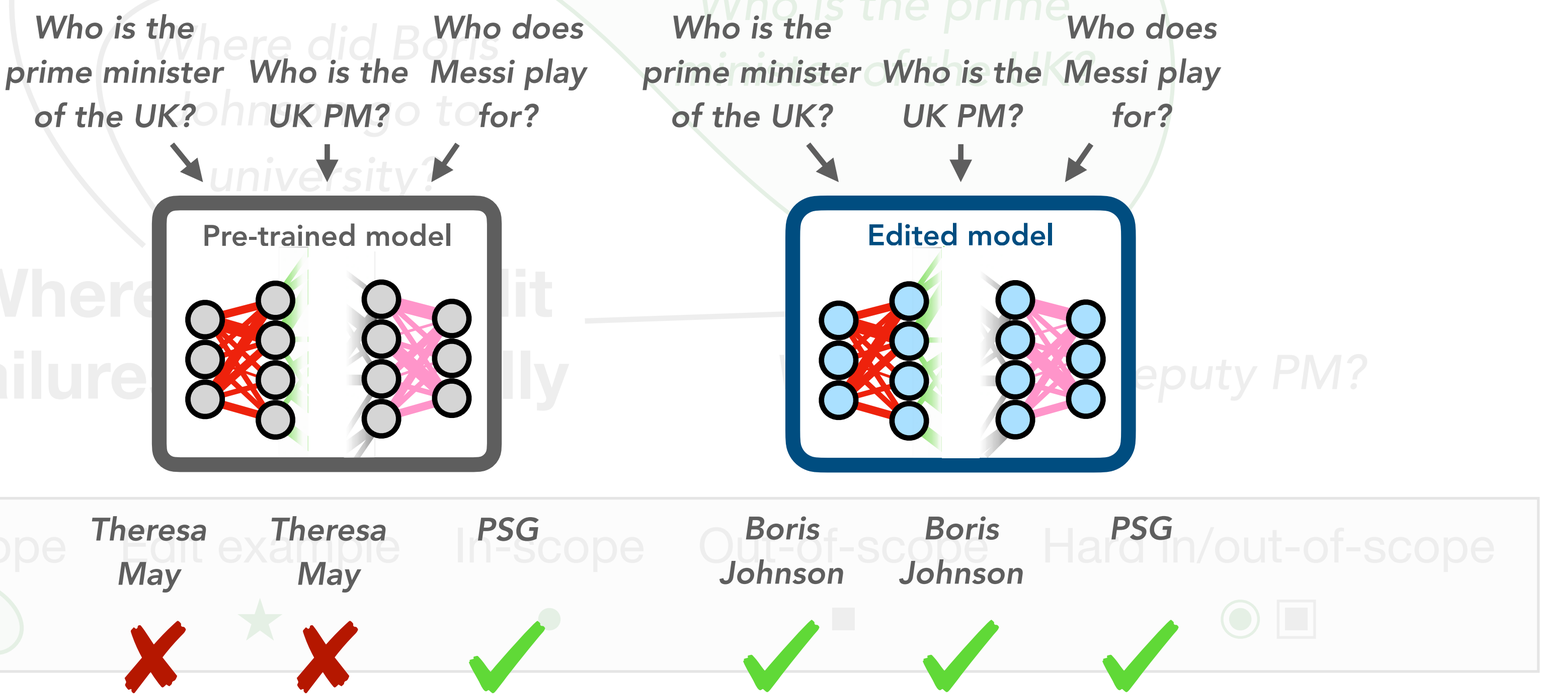
1. **Edit Success (ES):** ↑ accuracy on <u>in-scope</u> examples
2. **Drawdown (DD):** ↓ accuracy drop on <u>out-of-scope</u> examples



Edit *what*, exactly?

Metrics for evaluating model edits

1. **Edit Success (ES):** ↑ accuracy on <u>in-scope</u> examples
2. **Drawdown (DD):** ↓ accuracy drop on <u>out-of-scope</u> examples



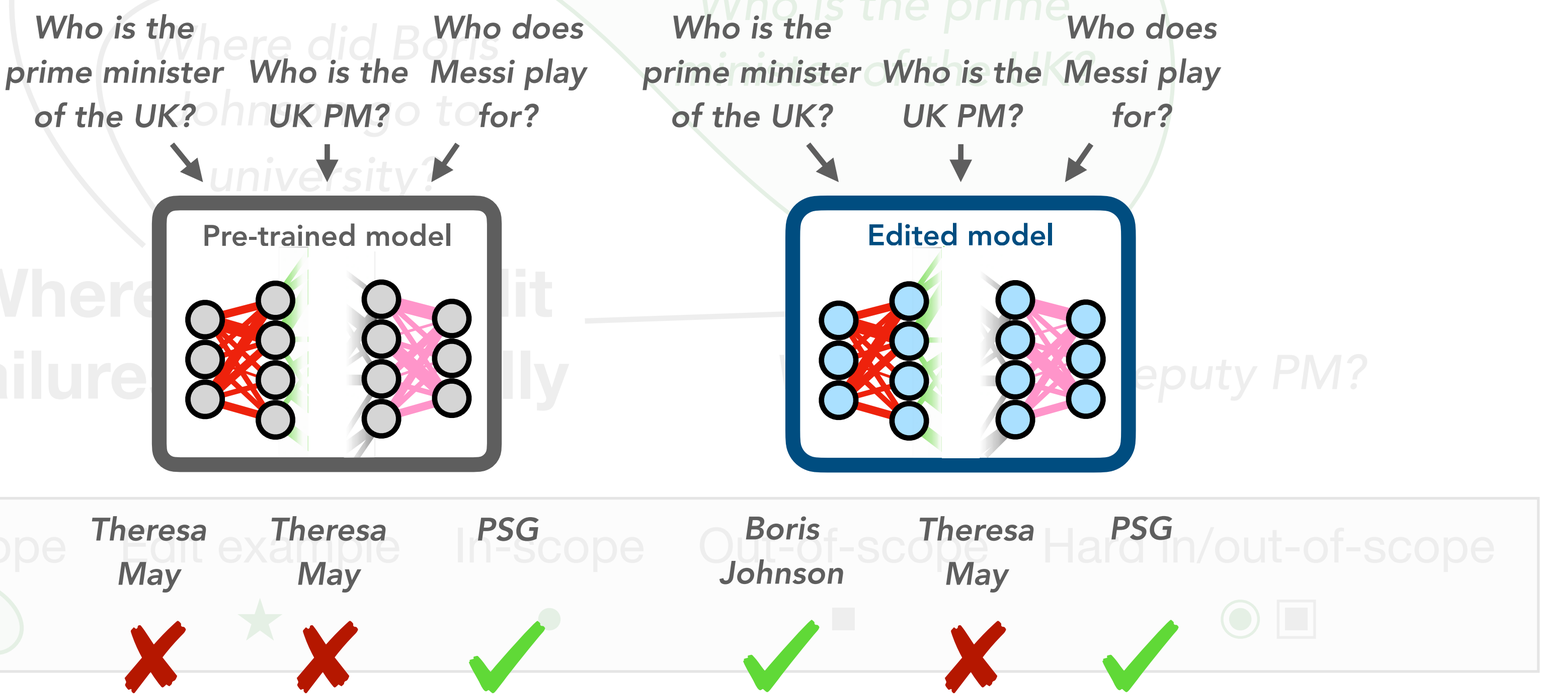
ES: 1

DD: 0

Edit *what*, exactly?

Metrics for evaluating model edits

1. **Edit Success (ES):** ↑ accuracy on <u>in-scope</u> examples
2. **Drawdown (DD):** ↓ accuracy drop on <u>out-of-scope</u> examples



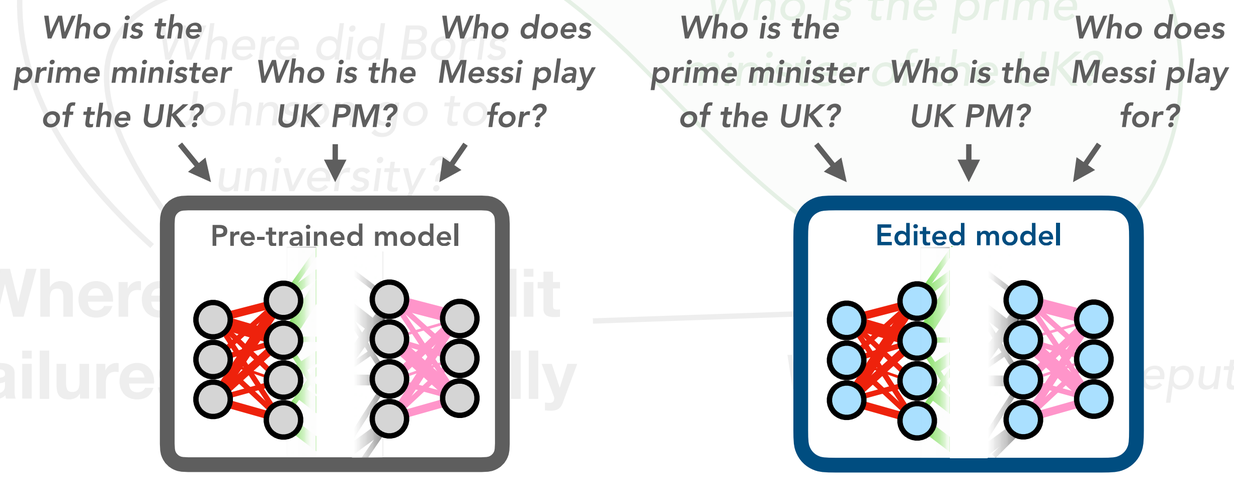
ES: 0.5

DD: 0

Edit *what*, exactly?

Metrics for evaluating model edits

1. **Edit Success (ES):** ↑ accuracy on <u>in-scope</u> examples
2. **Drawdown (DD):** ↓ accuracy drop on <u>out-of-scope</u> examples



ES: 1

DD: 1

Today's Plan

I. Background

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions

Existing approaches to editing

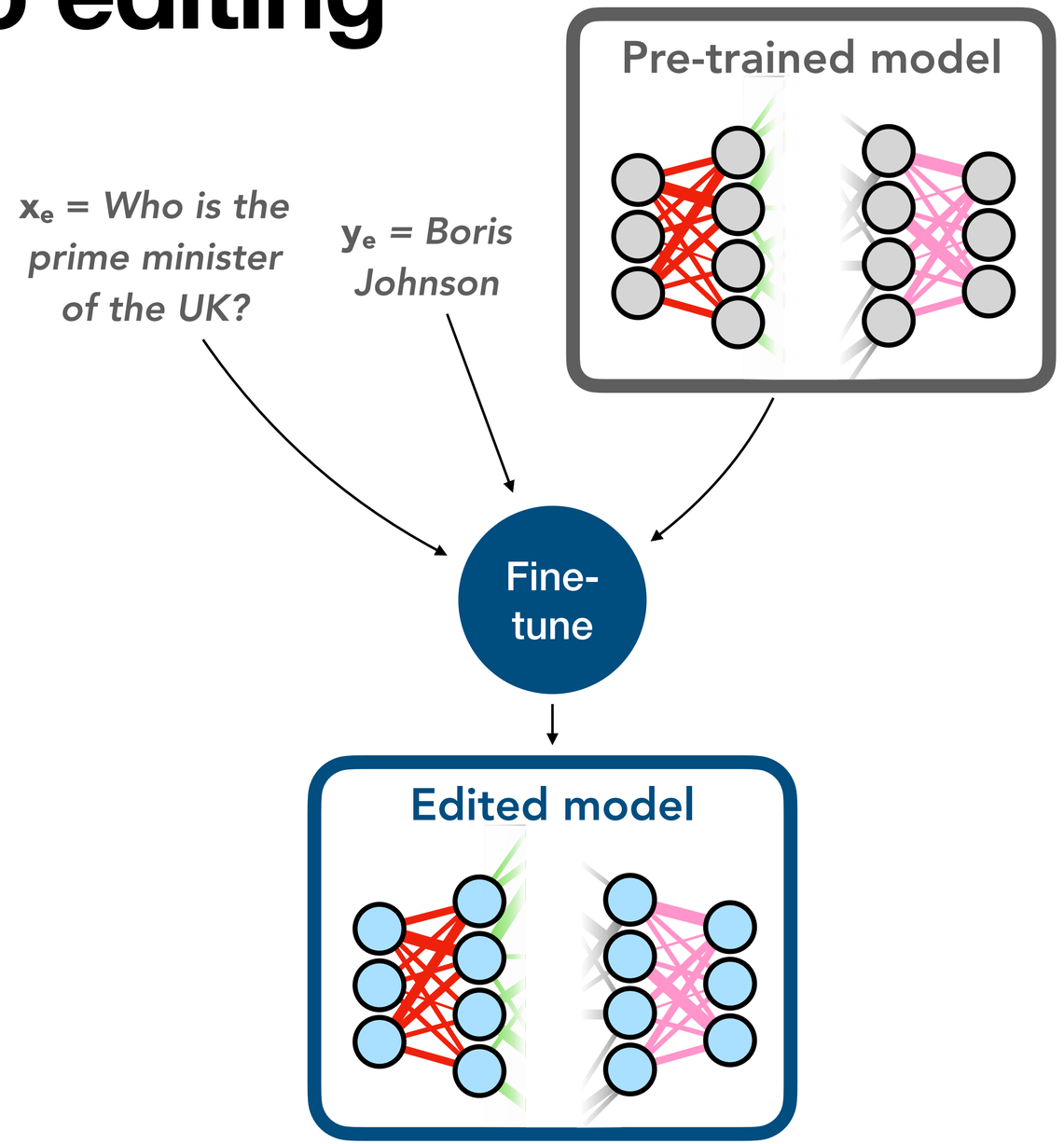
Some simple baselines

What about just fine-tuning?

Existing approaches to editing

Some simple baselines

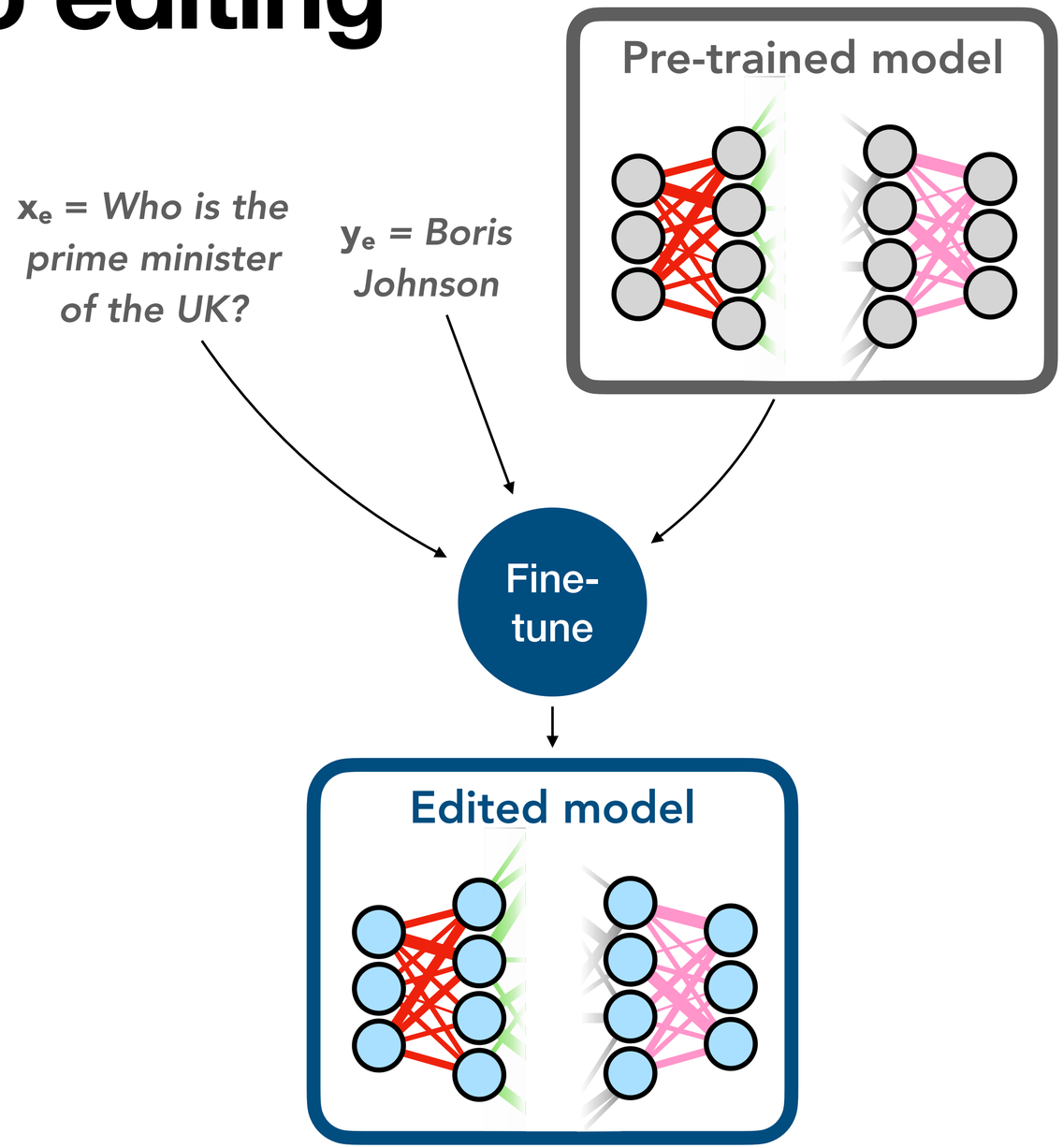
What about just fine-tuning?



Existing approaches to editing

Some simple baselines

What about just fine-tuning?

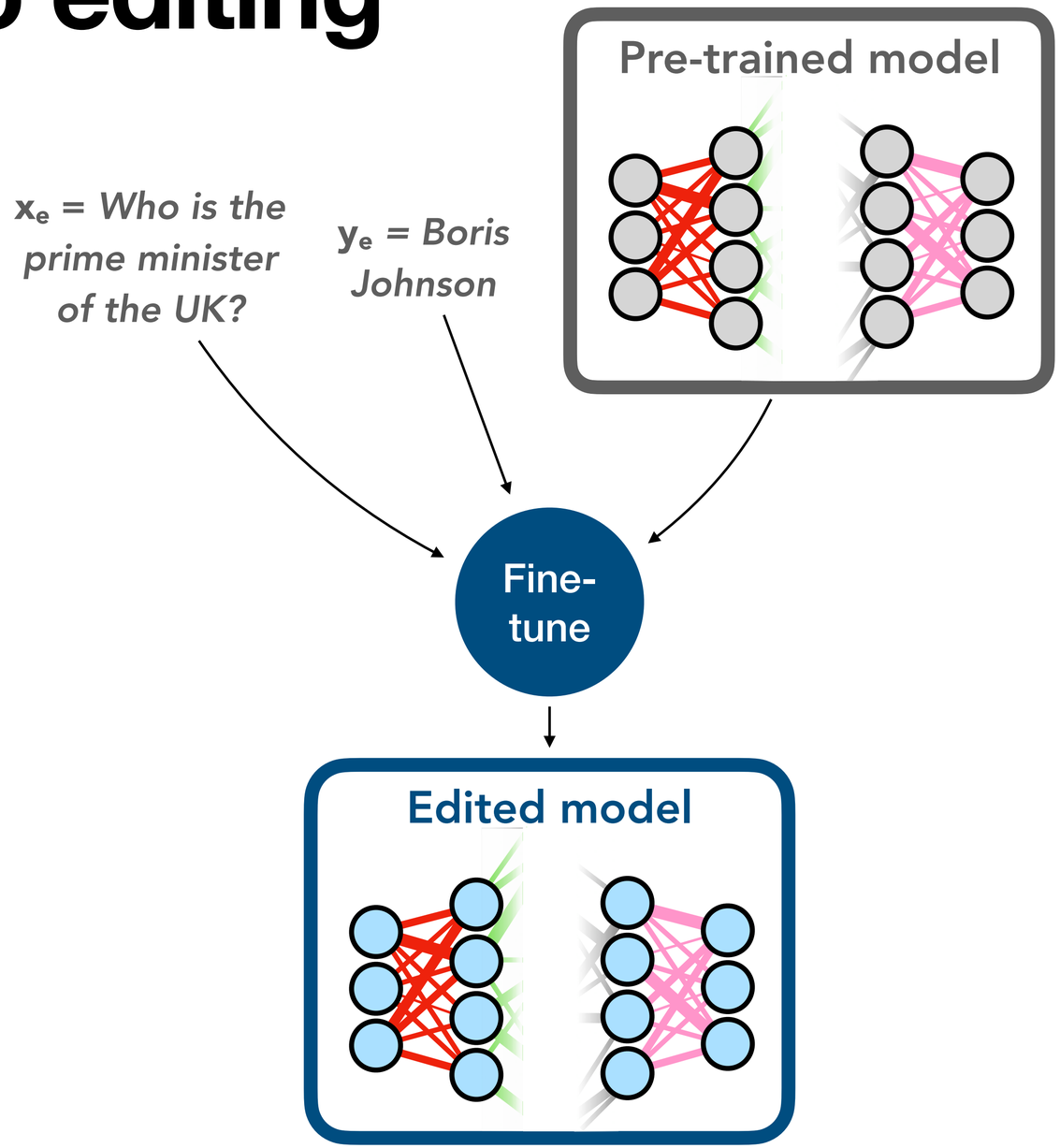


+ simple, universal

Existing approaches to editing

Some simple baselines

What about just fine-tuning?



+ simple, universal

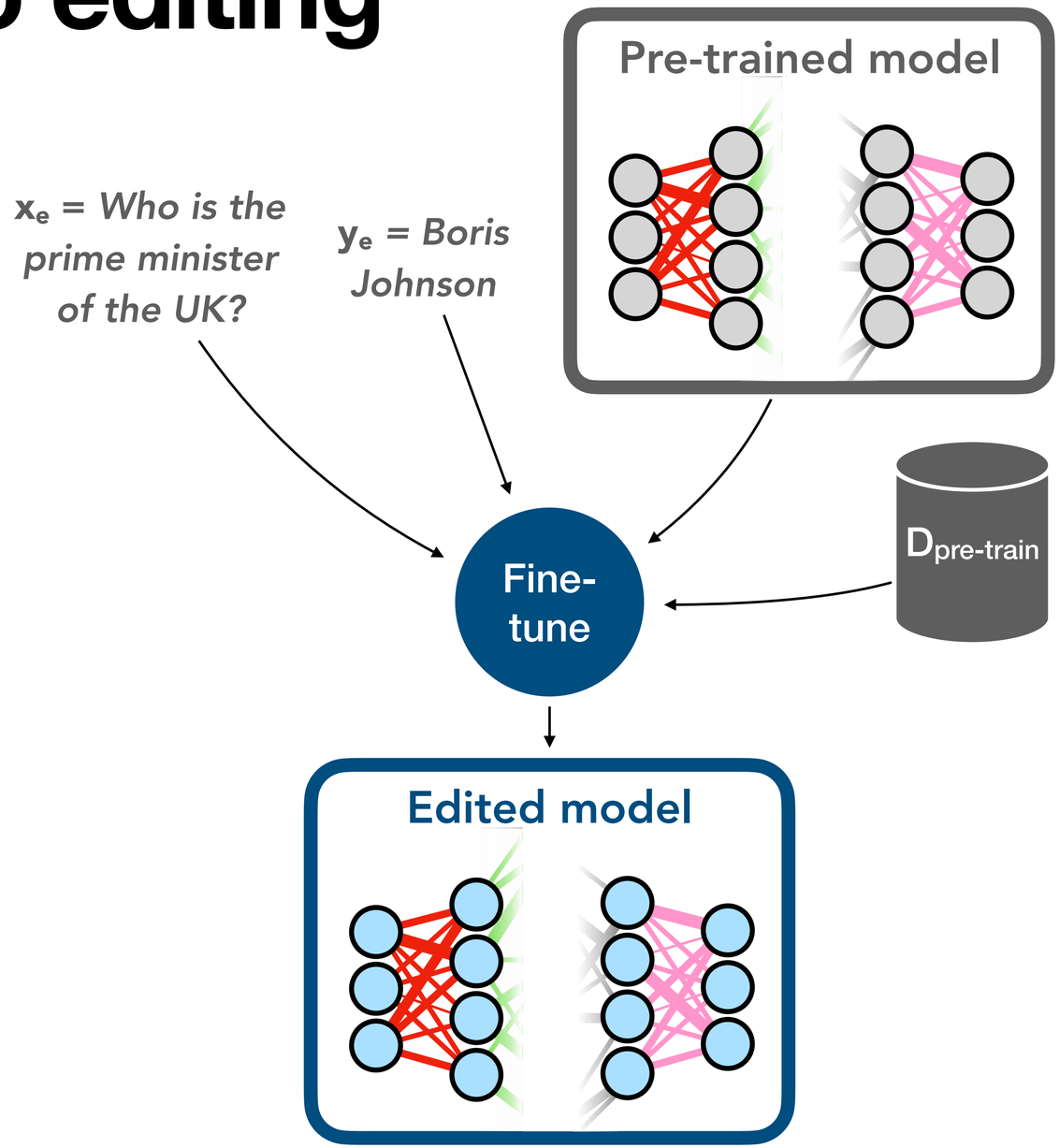
- undergeneralizes, overgeneralizes

Existing approaches to editing

Some simple baselines

What about just fine-tuning?

What if we add some training data?



+ simple, universal

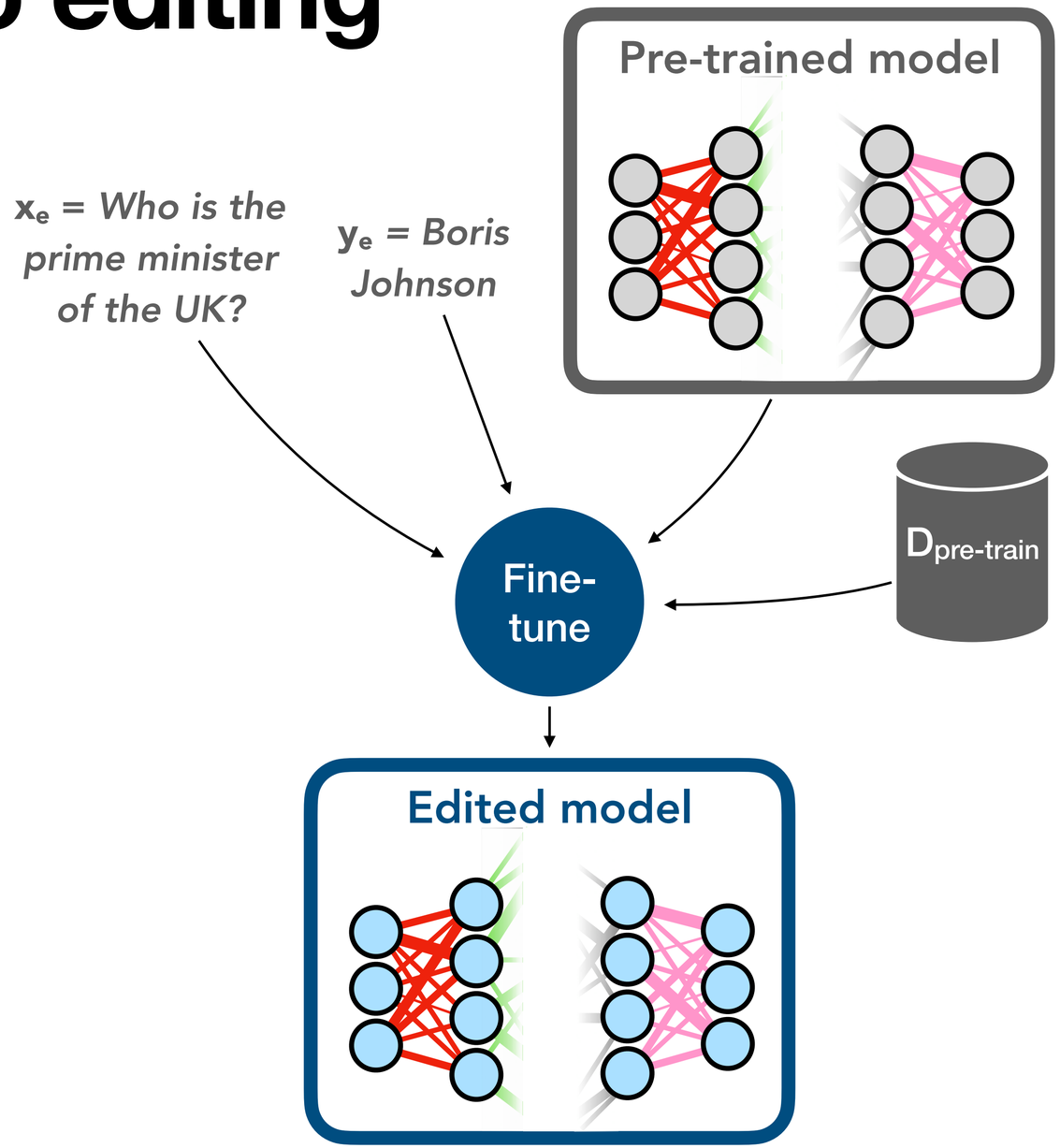
- undergeneralizes, overgeneralizes

Existing approaches to editing

Some simple baselines

What about just fine-tuning?

What if we add some training data?



+ simple, universal

- undergeneralizes, ~~overgeneralizes~~

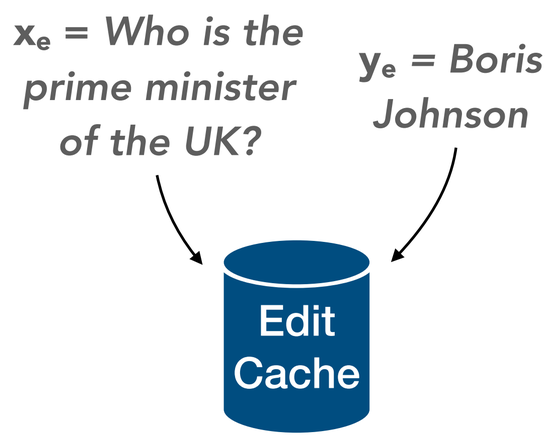
Existing approaches to editing

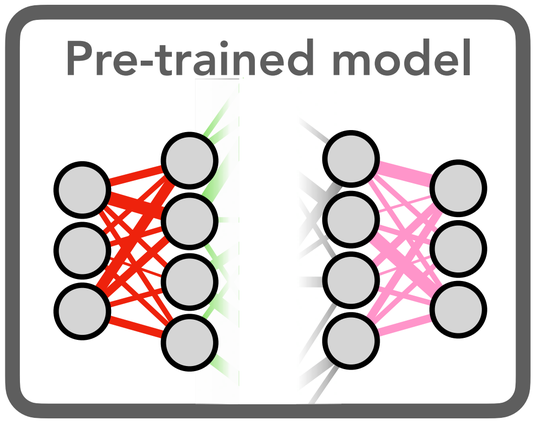
Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits





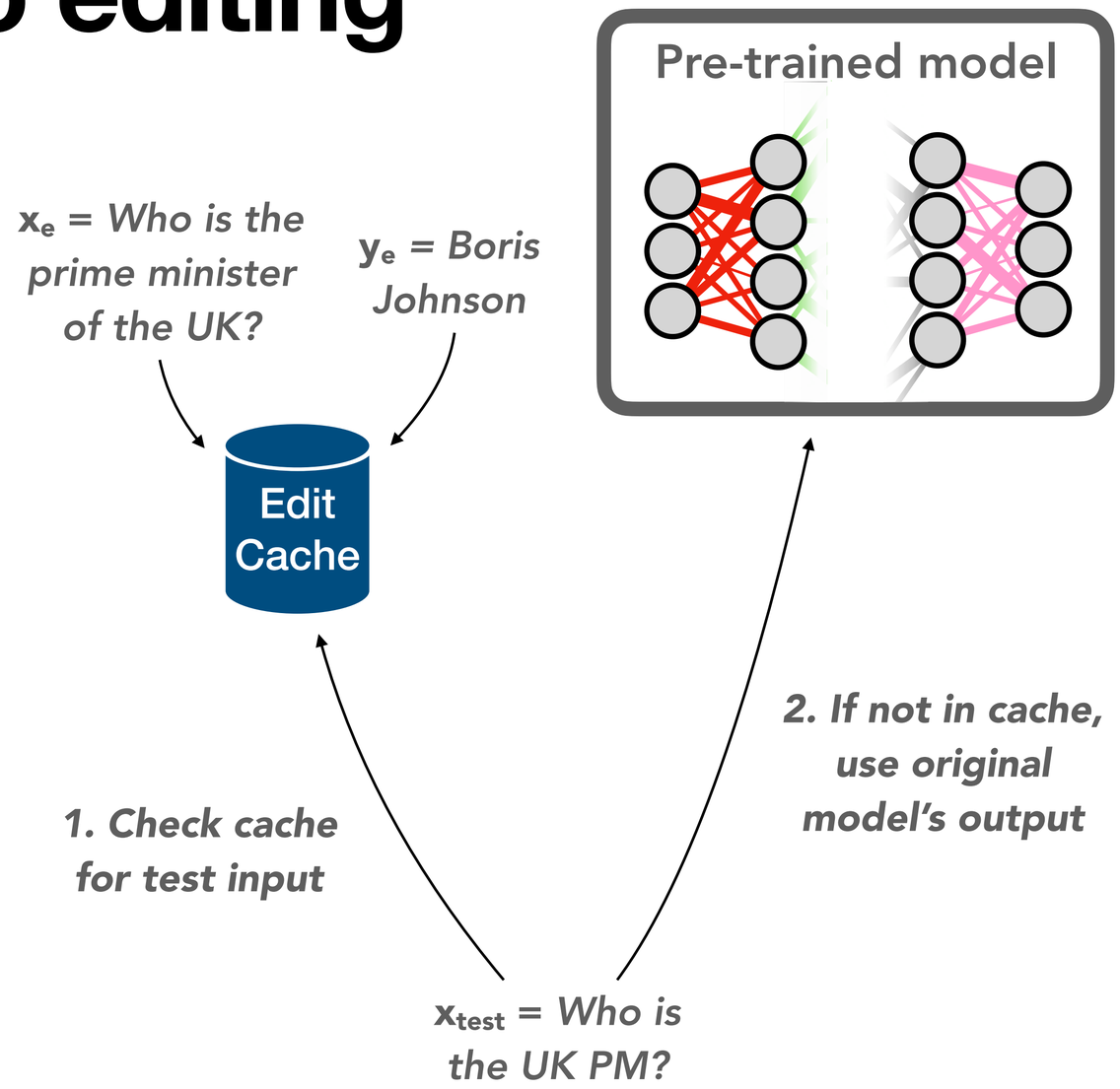
Existing approaches to editing

Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits



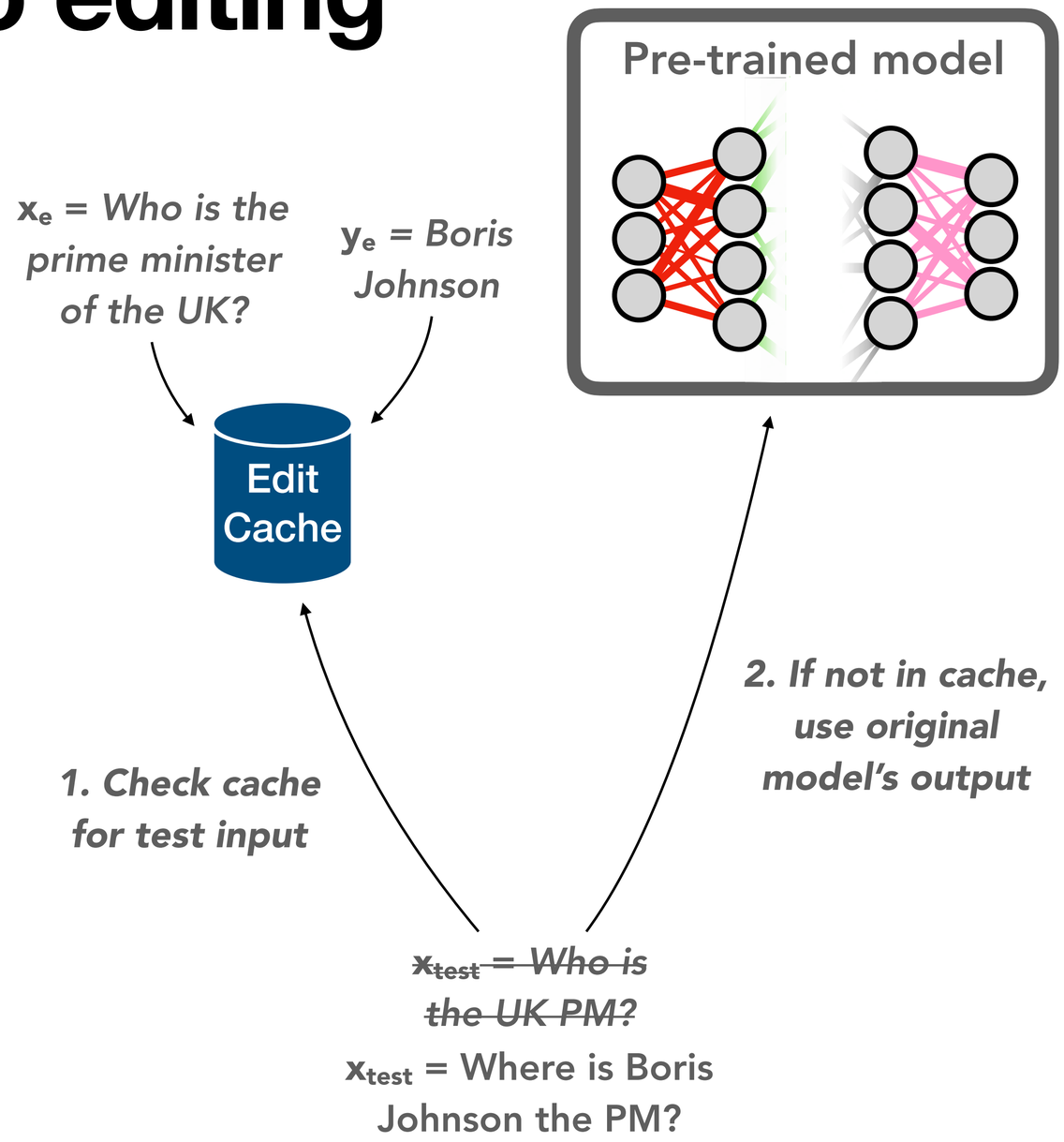
Existing approaches to editing

Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits



Existing approaches to editing

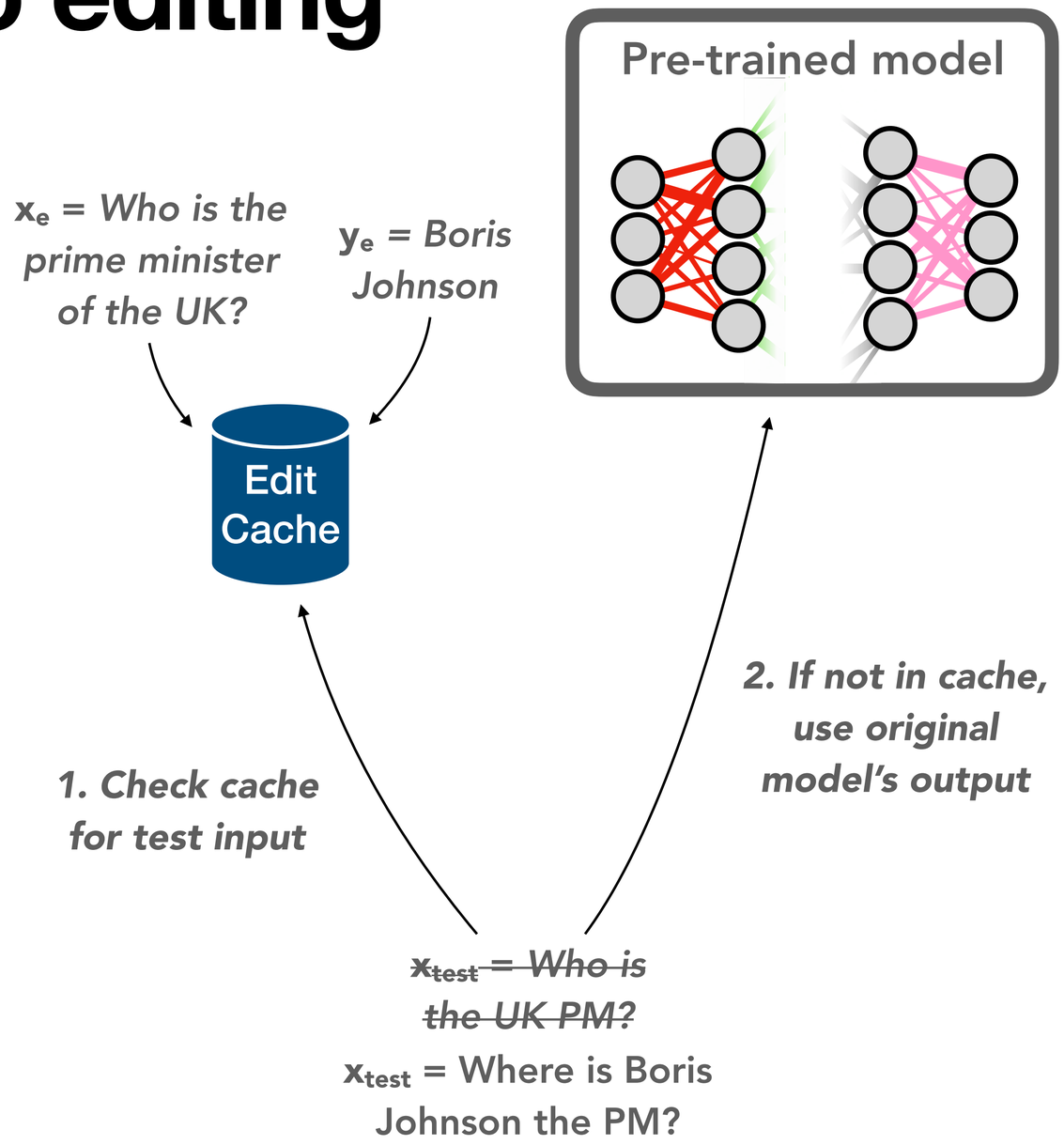
Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits

Can we **learn** a more expressive edit rule from data?



Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, \quad x_{loc}, \quad x_{in}, \quad y_{in}) \}$

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, \quad x_{loc}, \quad x_{in}, \quad y_{in}) \}$

z_{edit} = “Who is the UK PM? Boris Johnson”

x_{loc} = “What team does Messi play for?”

x_{in} = “The prime minister of the UK is currently who?”

y_{in} = “Boris Johnson”

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

Enforce locality with out-of-scope example

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

Enforce locality with out-of-scope example

Enforce generalization with in-scope example

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

Enforce locality with out-of-scope example

Enforce generalization with in-scope example

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_{\phi}(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

Enforce locality with out-of-scope example

Enforce generalization with in-scope example

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_{\phi}(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

Perform edit

Enforce locality with out-of-scope example

Enforce generalization with in-scope example

z_{edit} = "Who is the UK PM? Boris Johnson"

x_{loc} = "What team does Messi play for?"

x_{in} = "The prime minister of the UK is currently who?"

y_{in} = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_{\phi}(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Outer loop

(check if edit worked)

$$L_{edit} = p_{\theta'}(\mathbf{y_{in}} \mid \mathbf{x_{in}})$$

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $\mathbf{D_{edit}} = \{ (\mathbf{z_{edit}}, \mathbf{x_{loc}}, \mathbf{x_{in}}, \mathbf{y_{in}}) \}$

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example

$\mathbf{z_{edit}}$ = "Who is the UK PM? Boris Johnson"

$\mathbf{x_{loc}}$ = "What team does Messi play for?"

$\mathbf{x_{in}}$ = "The prime minister of the UK is currently who?"

$\mathbf{y_{in}}$ = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_\phi(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Outer loop

(check if edit worked)

$$L_{edit} = p_{\theta'}(\mathbf{y_{in}} \mid \mathbf{x_{in}})$$

Did predictions **change** **where** we wanted them to?

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $\mathbf{D_{edit}} = \{ (\mathbf{z_{edit}}, \mathbf{x_{loc}}, \mathbf{x_{in}}, \mathbf{y_{in}}) \}$

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example

$\mathbf{z_{edit}}$ = "Who is the UK PM? Boris Johnson"

$\mathbf{x_{loc}}$ = "What team does Messi play for?"

$\mathbf{x_{in}}$ = "The prime minister of the UK is currently who?"

$\mathbf{y_{in}}$ = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_\phi(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Outer loop

(check if edit worked)

$$L_{edit} = p_{\theta'}(\mathbf{y_{in}} \mid \mathbf{x_{in}})$$

$$L_{loc} = \text{KL}\left(p_\theta(\cdot \mid \mathbf{x_{loc}}) \,\|\, p_{\theta'}(\cdot \mid \mathbf{x_{loc}}) \right)$$

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $\mathbf{D_{edit}} = \{ (\mathbf{z_{edit}}, \mathbf{x_{loc}}, \mathbf{x_{in}}, \mathbf{y_{in}}) \}$

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example

$\mathbf{z_{edit}}$ = "Who is the UK PM? Boris Johnson"

$\mathbf{x_{loc}}$ = "What team does Messi play for?"

$\mathbf{x_{in}}$ = "The prime minister of the UK is currently who?"

$\mathbf{y_{in}}$ = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_\phi(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Outer loop

(check if edit worked)

$$L_{edit} = p_{\theta'}(\mathbf{y_{in}} \mid \mathbf{x_{in}})$$

Did we keep predictions **the same** everywhere else?

$$L_{loc} = \text{KL}\left(p_\theta(\cdot \mid \mathbf{x_{loc}}) \,\|\, p_{\theta'}(\cdot \mid \mathbf{x_{loc}}) \right)$$

Learning to edit

Editing as meta-learning

Requirement: an "edit dataset" $\mathbf{D_{edit}} = \{ (\mathbf{z_{edit}}, \mathbf{x_{loc}}, \mathbf{x_{in}}, \mathbf{y_{in}}) \}$

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example

$\mathbf{z_{edit}}$ = "Who is the UK PM? Boris Johnson"

$\mathbf{x_{loc}}$ = "What team does Messi play for?"

$\mathbf{x_{in}}$ = "The prime minister of the UK is currently who?"

$\mathbf{y_{in}}$ = "Boris Johnson"

Inner loop

(run the editor)

$$\theta' = \text{Edit}_\phi(\theta, \mathbf{z_{edit}})$$

(optional) editor parameters

Outer loop

(check if edit worked)

$$L_{edit} = p_{\theta'}(\mathbf{y_{in}} \mid \mathbf{x_{in}})$$

$$L_{loc} = \text{KL}\left(p_\theta(\cdot \mid \mathbf{x_{loc}}) \,\|\, p_{\theta'}(\cdot \mid \mathbf{x_{loc}}) \right)$$

Backprop $L_{edit} + L_{loc}$ back into base model/editor

Learning to edit

A tale of two meta-learning frameworks

1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit

Learning to edit

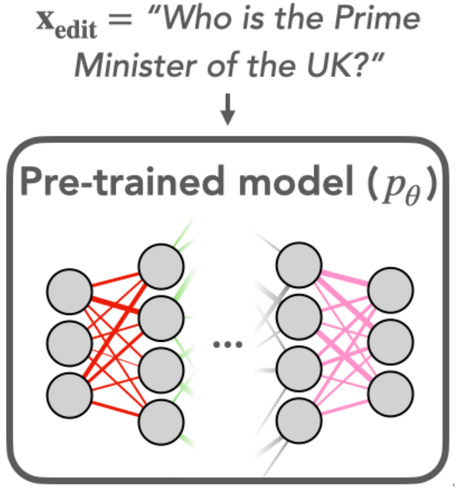
A tale of two meta-learning frameworks

1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta \, p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit
2. Hypernetwork-based: Freeze base model, train a **gradient transform** $g_\phi(\,\cdot\,)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\,\nabla_\theta \, p_\theta(\mathbf{z}_{\mathbf{edit}}))$ gives a good edit

Learning to edit

A tale of two meta-learning frameworks

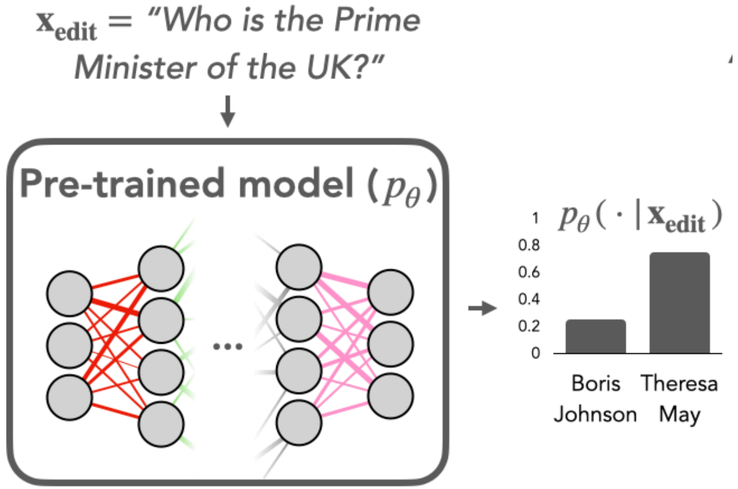
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z_{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

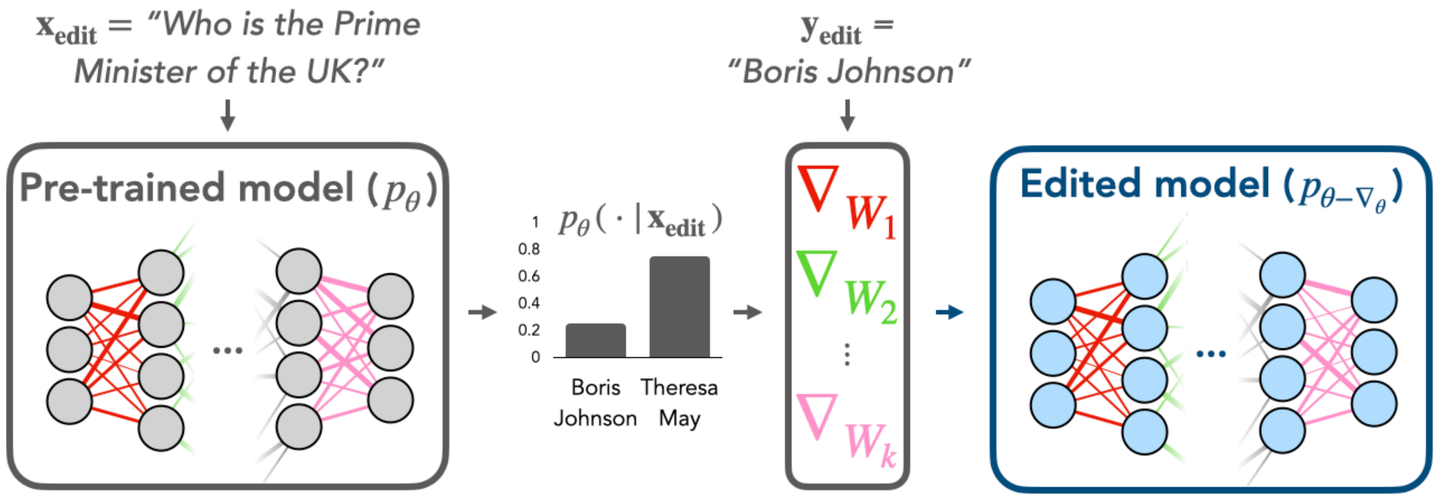
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

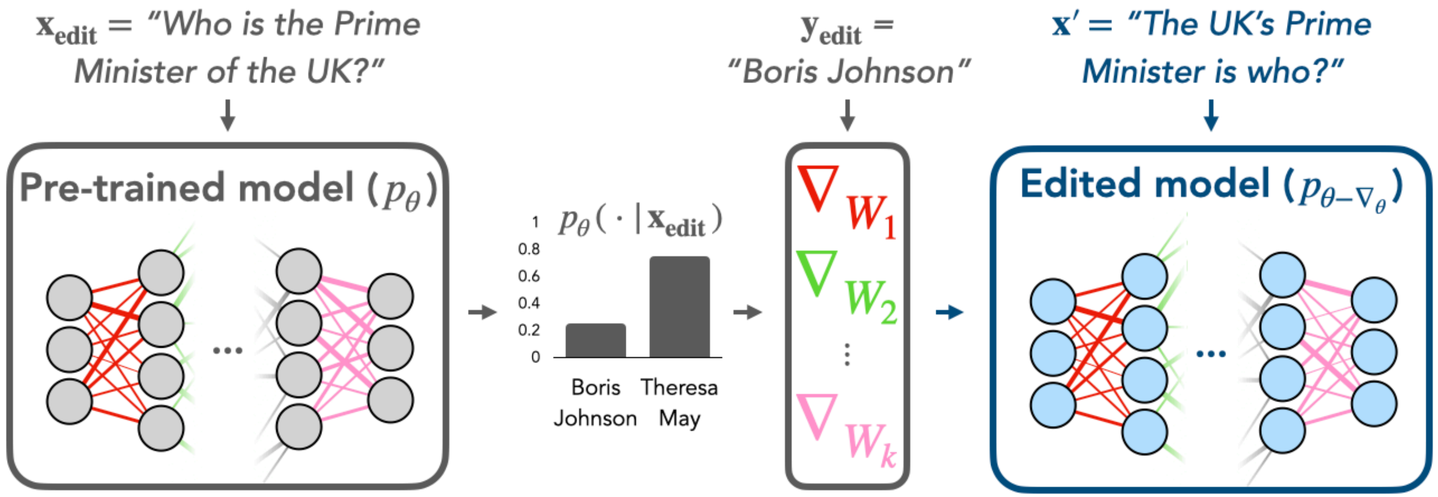
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

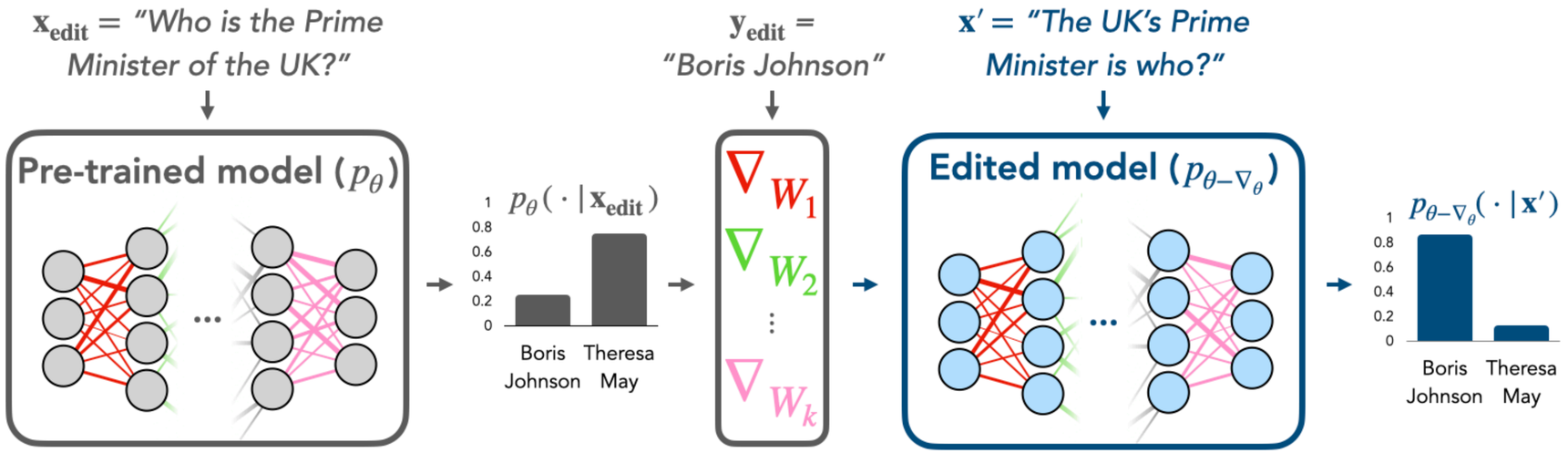
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

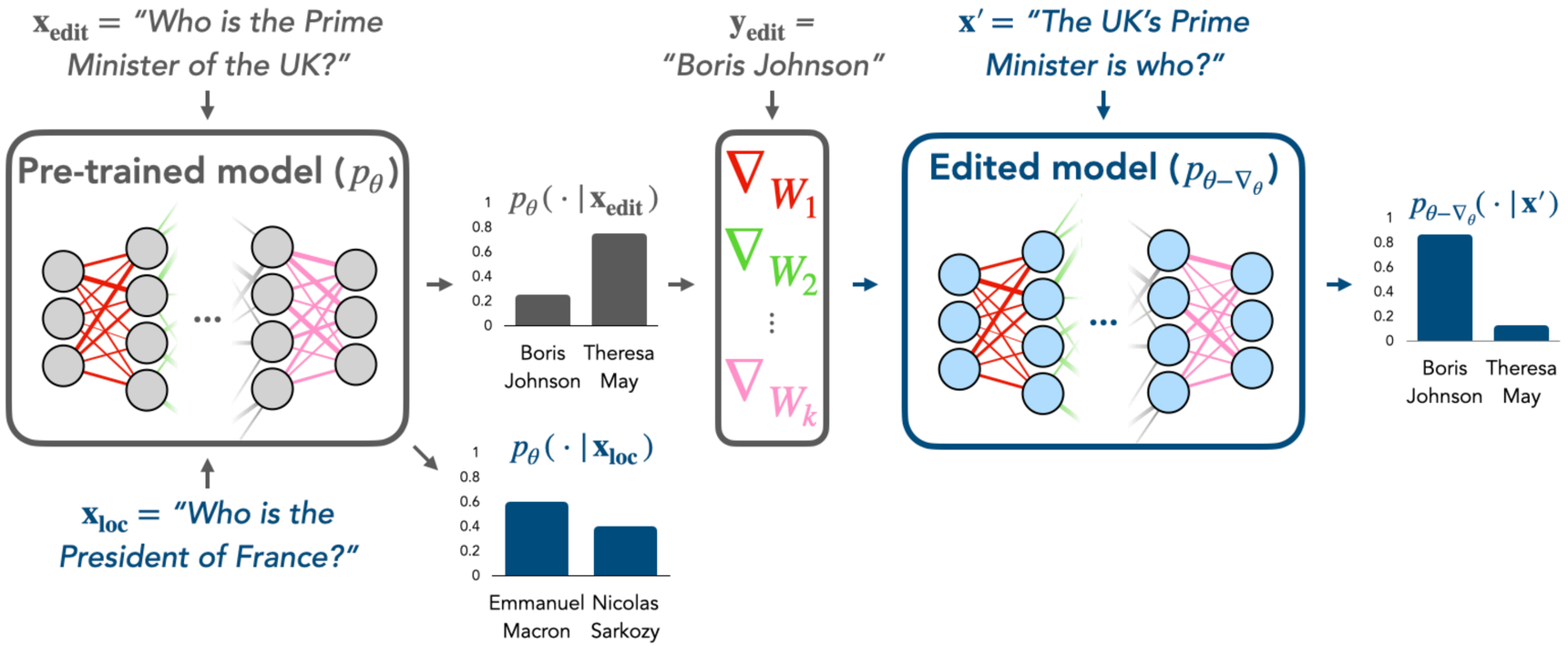
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

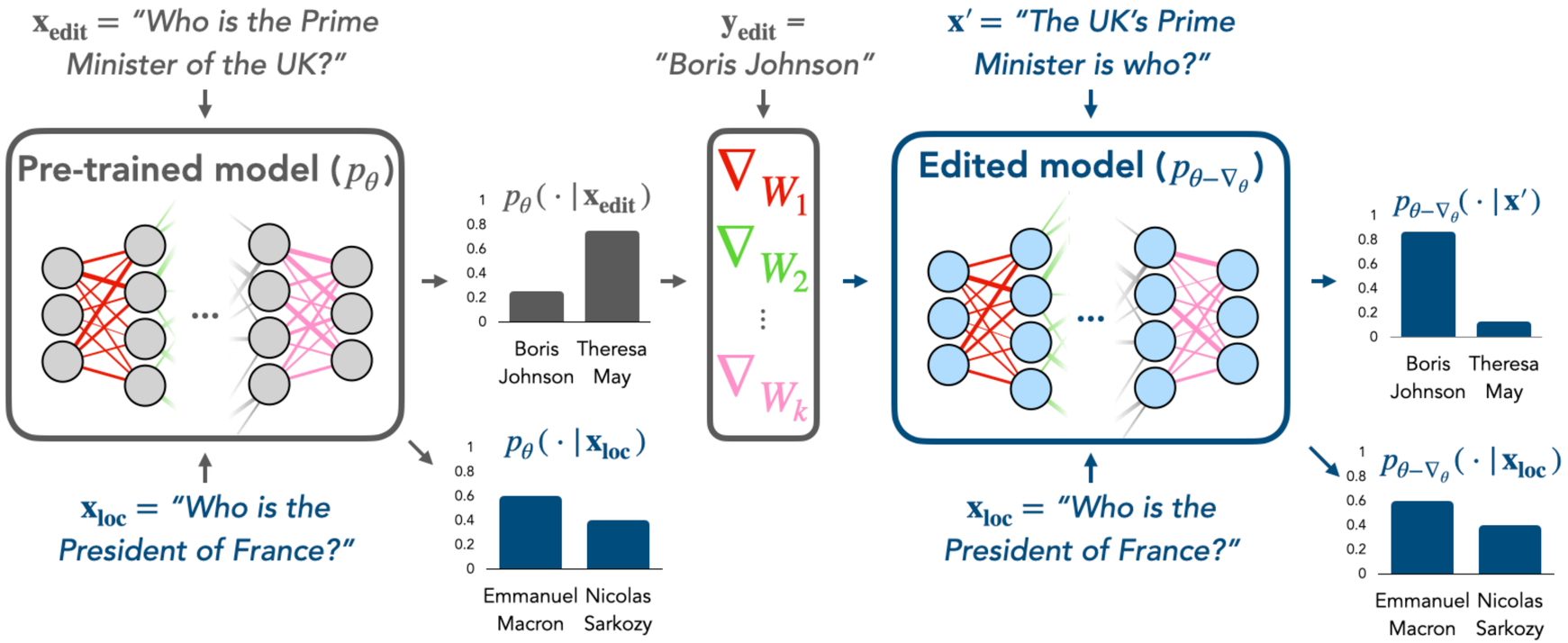
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

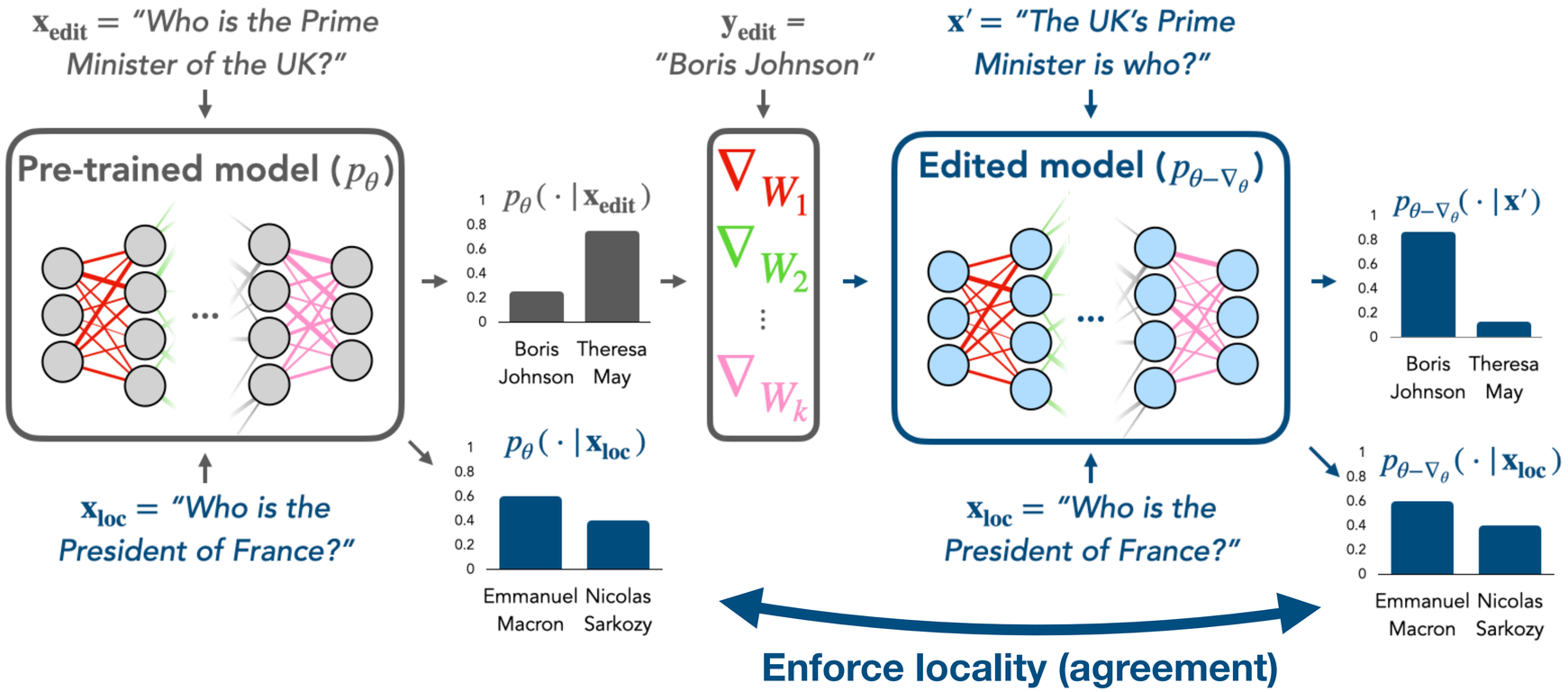
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

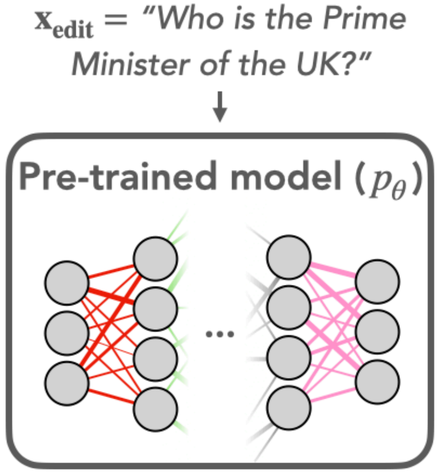
1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(\mathbf{z}_{\mathbf{edit}})$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

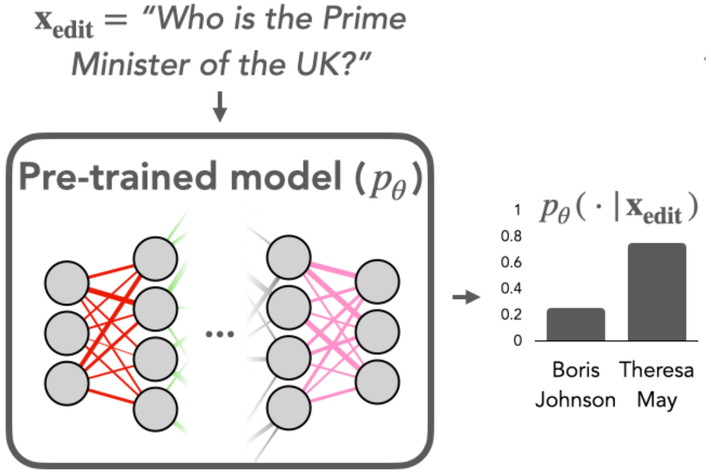
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z_{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

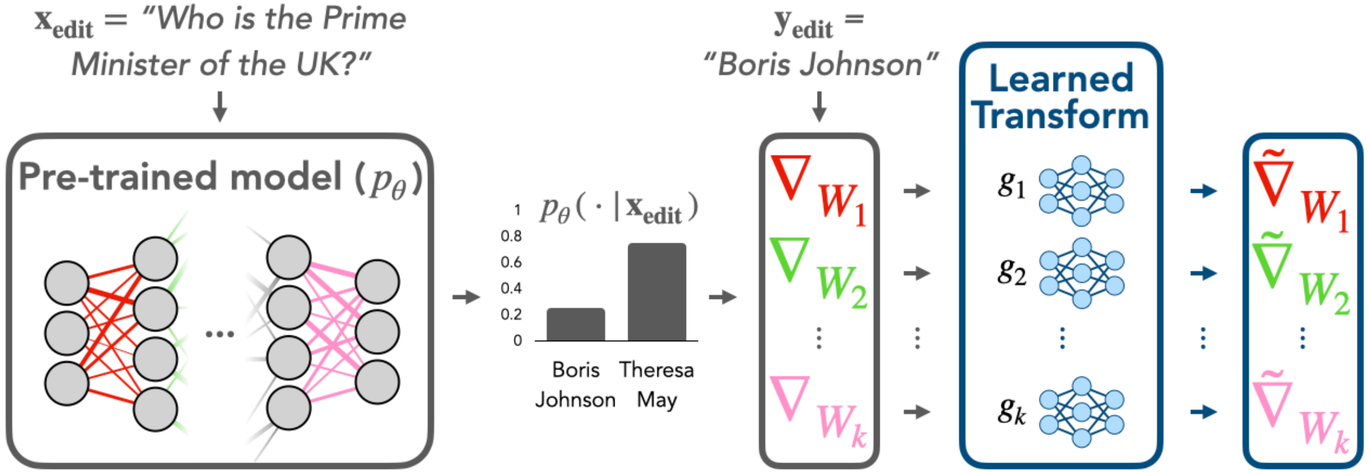
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z_{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

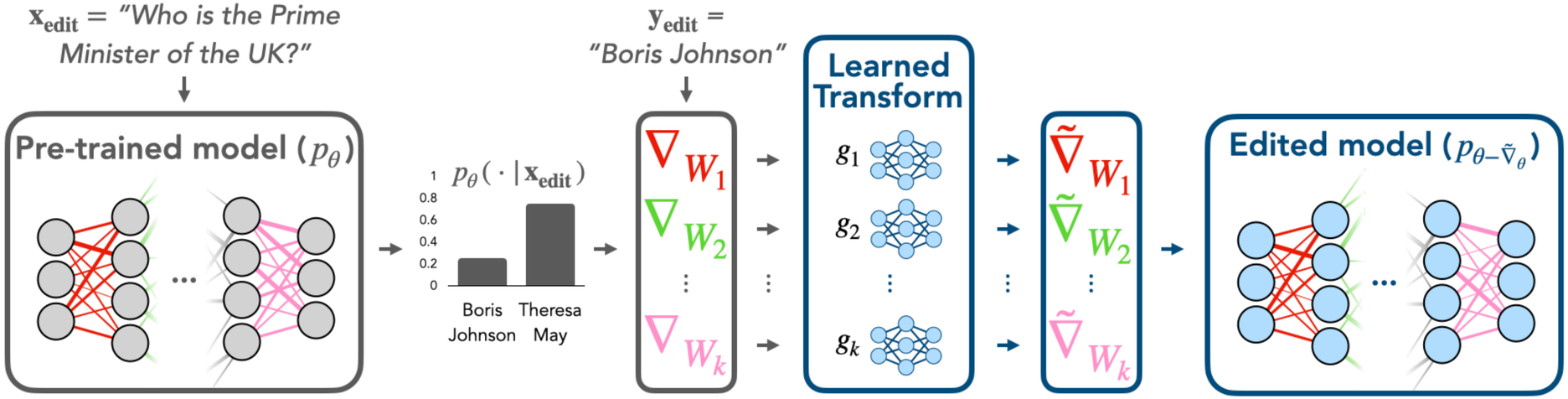
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z_{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

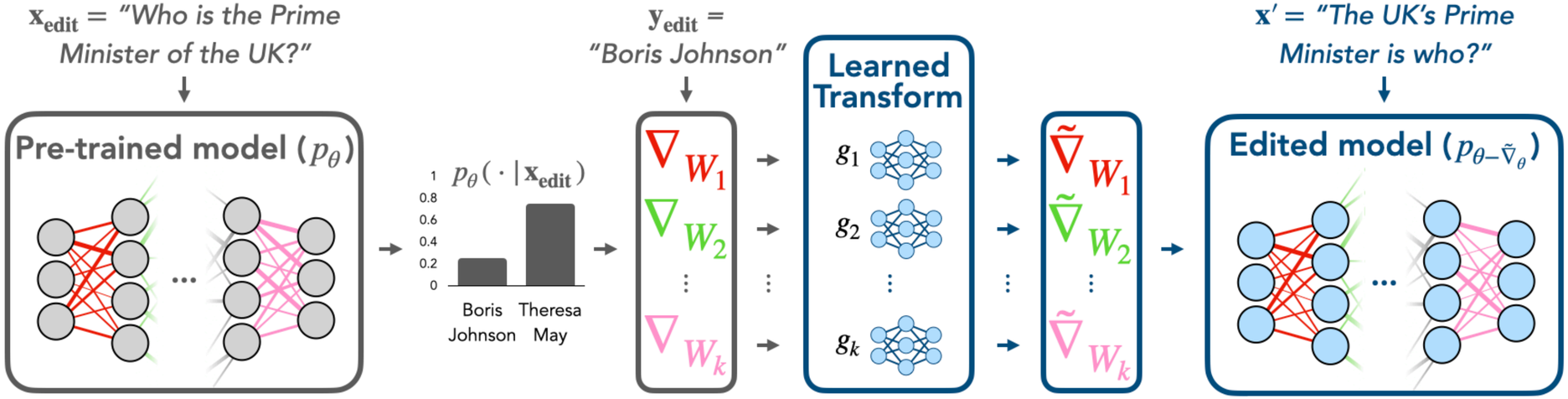
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z_{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

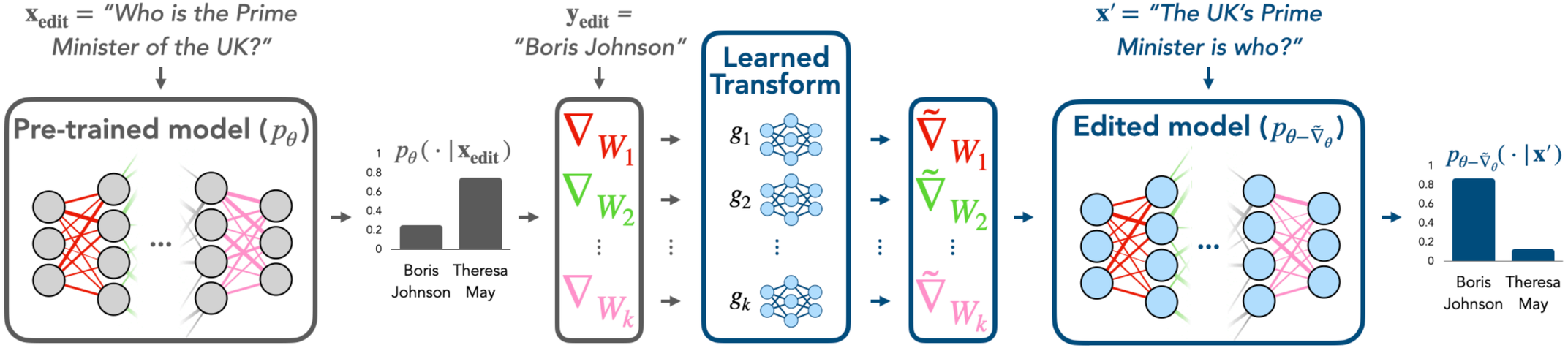
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

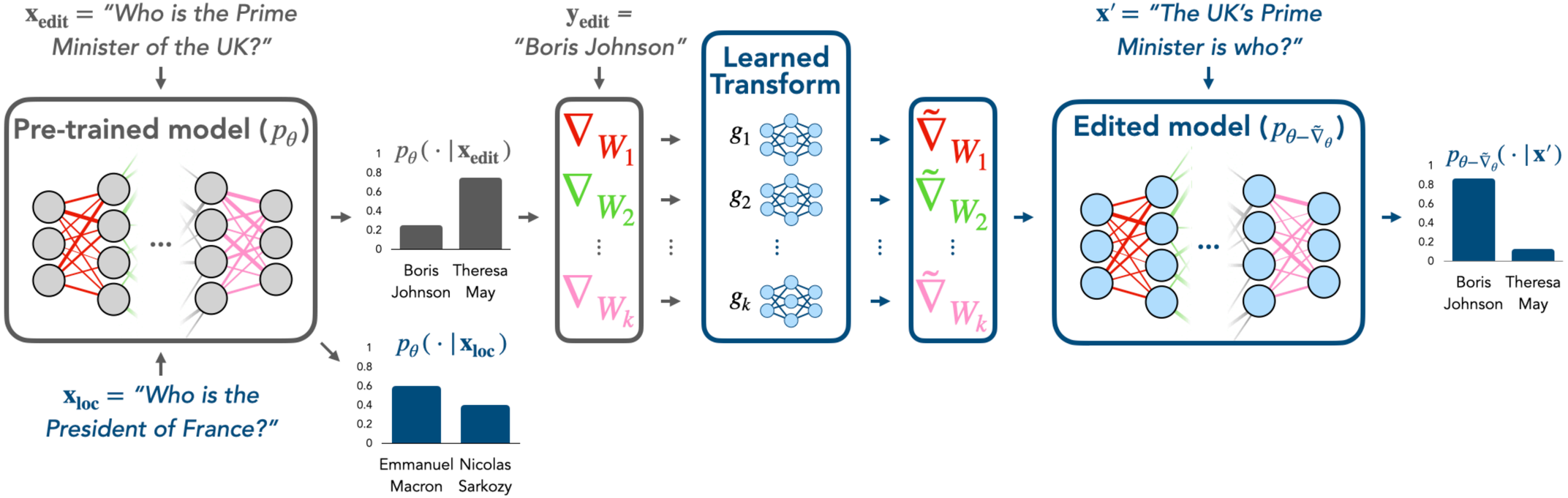
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

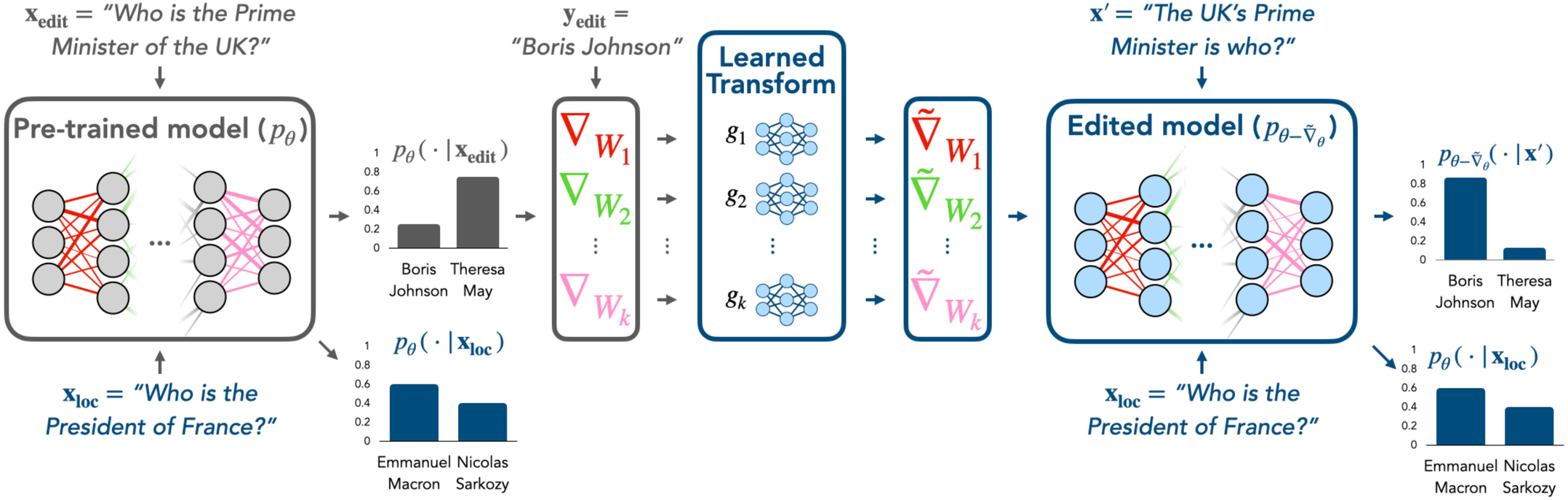
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}}))$ gives a good edit



Learning to edit

A tale of two meta-learning frameworks

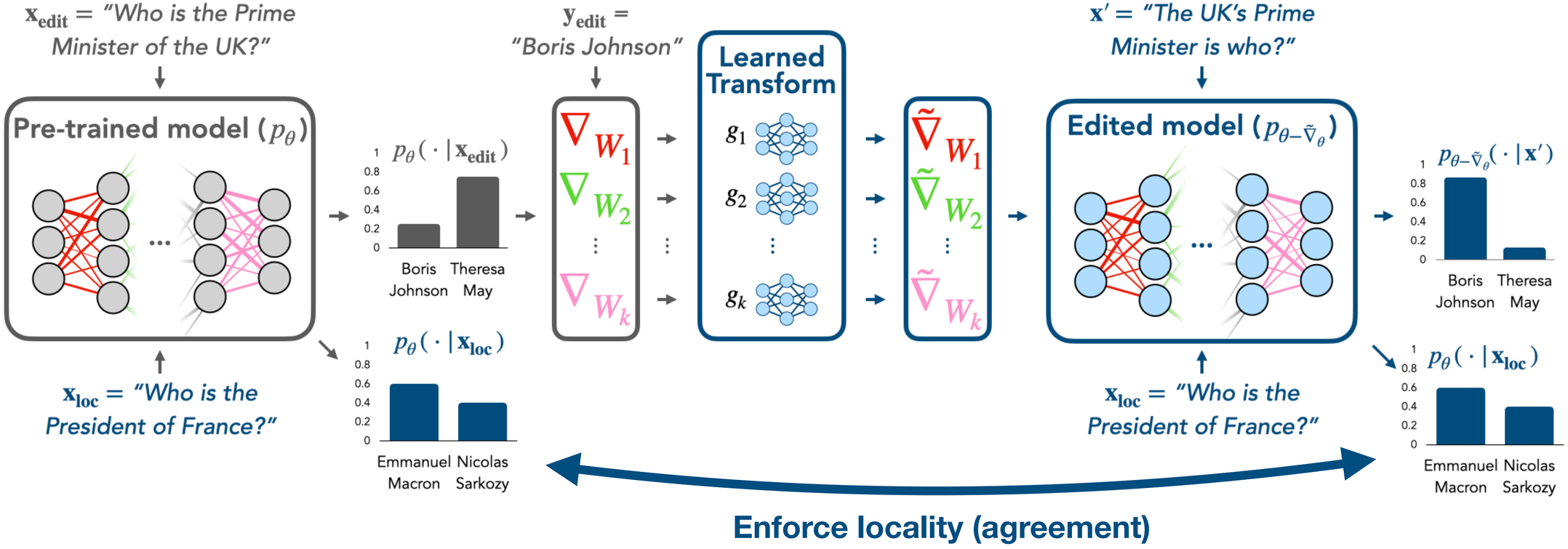
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}}))$ gives a good edit



Learning to edit

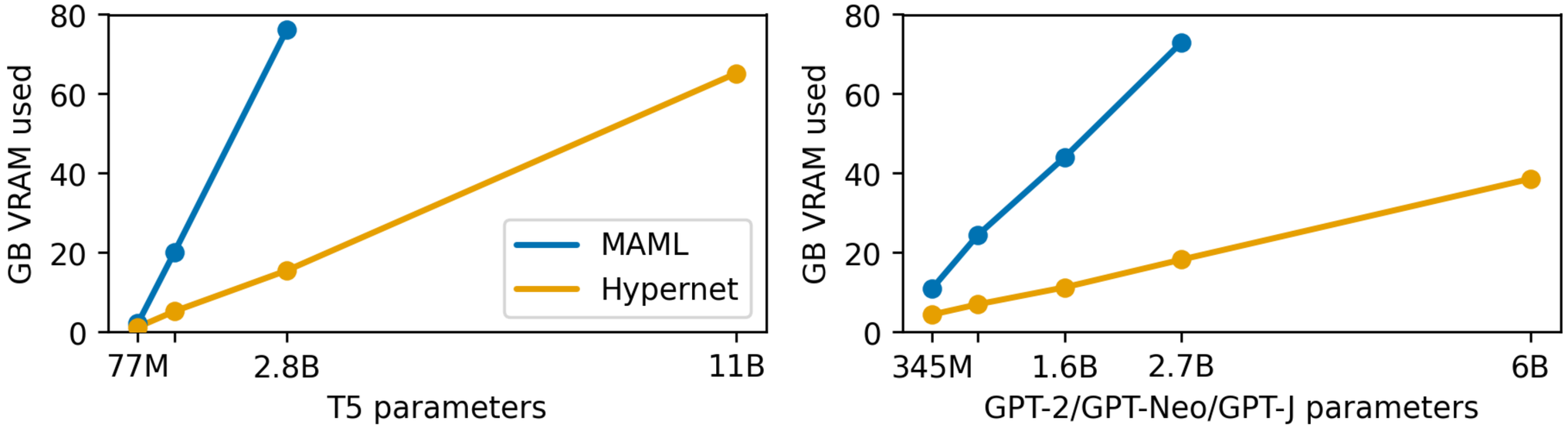
A tale of two meta-learning frameworks

2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(\mathbf{z}_{\text{edit}}))$ gives a good edit



Challenges of editing at scale

- MAML-based editors show **high memory consumption**



Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

95% ES at 0.1B parameters

Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

95% ES at 0.1B parameters

(BART-base)

Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

95% ES at 0.1B parameters → 4% ES at 11B parameters

(BART-base)

Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

95% ES at 0.1B parameters → 4% ES at 11B parameters

(BART-base) (T5-XXL)

Challenges of editing at scale

- MAML-based editors show **high memory consumption**
- Hypernets use **restrictive approximations** to handle high-dim parameters

95% ES at 0.1B parameters → 4% ES at 11B parameters

(BART-base) (T5-XXL)

Can we develop an **efficient** and **expressive** gradient transform?

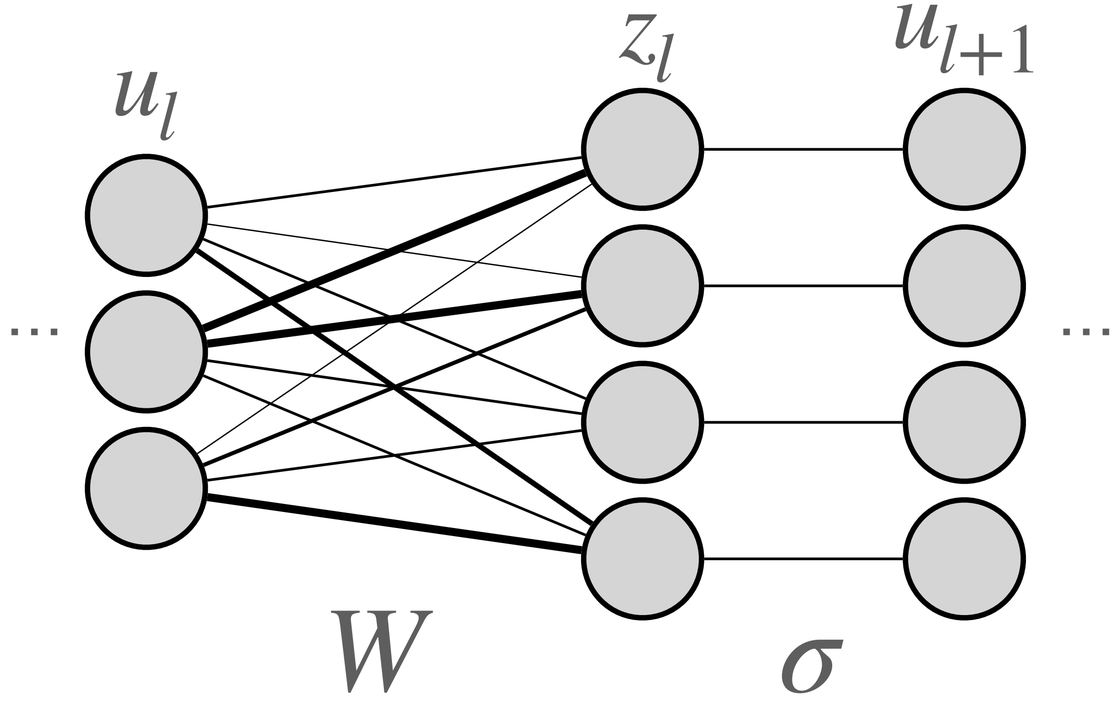
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

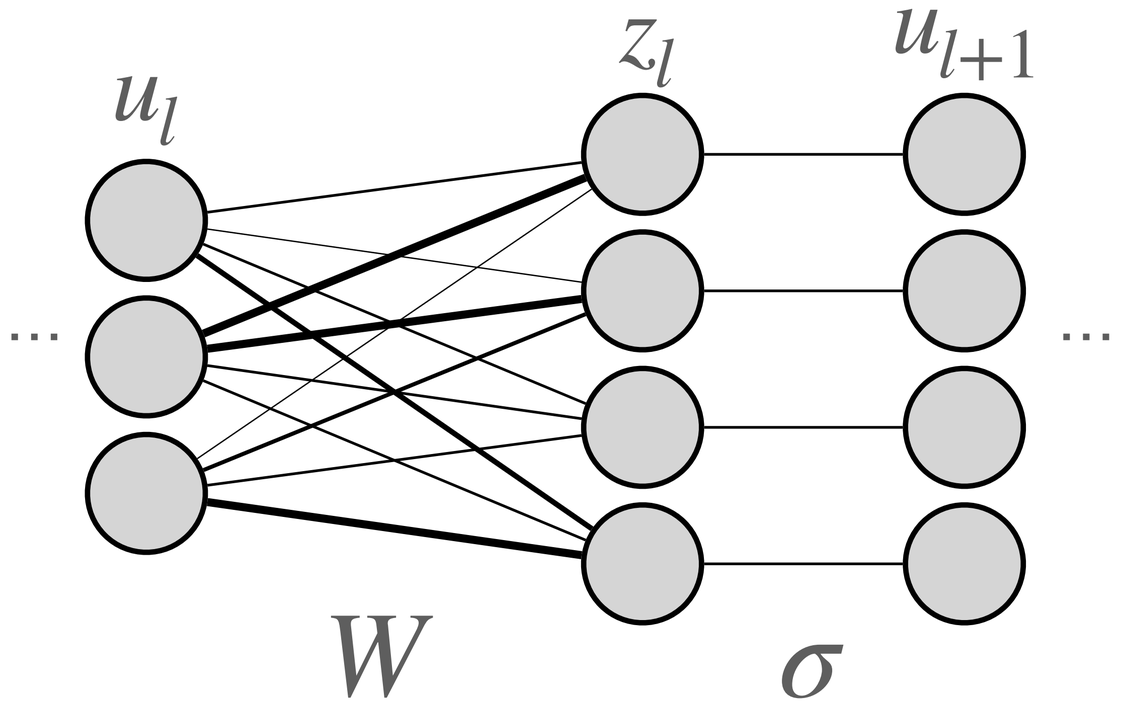


Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Forward pass computes u_l



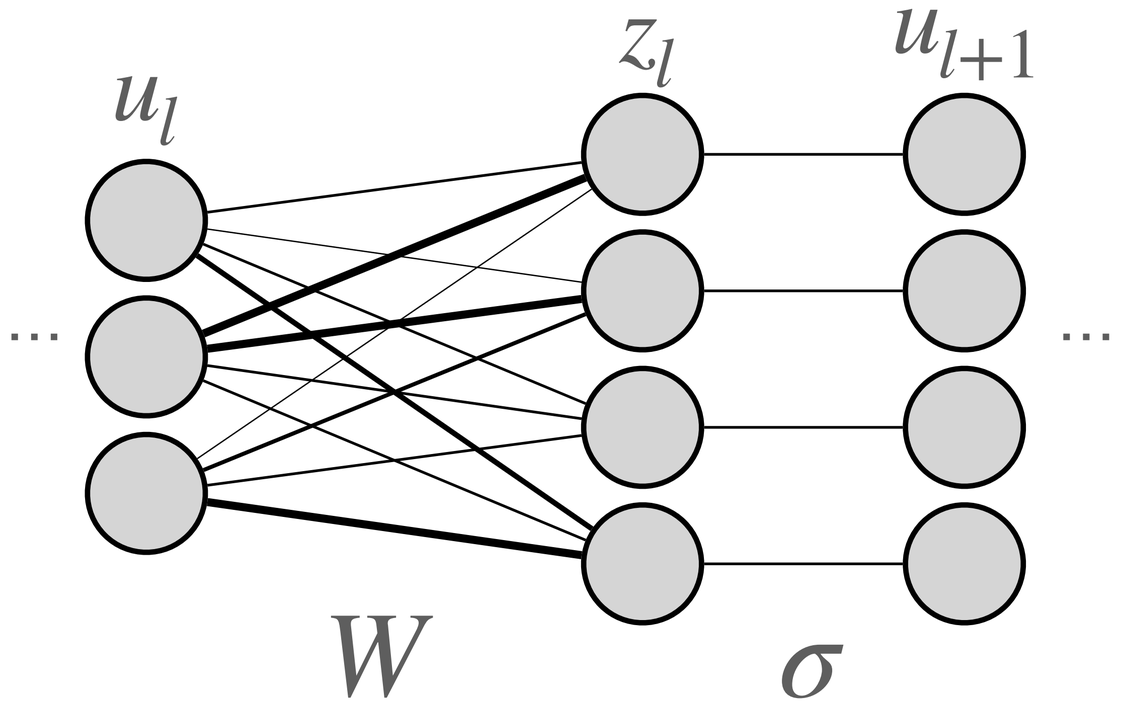
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Forward pass computes u_l

Backward pass computes $\delta_l = \dfrac{\partial L}{\partial z_l}$



Model Editor Networks using gradient Decomposition

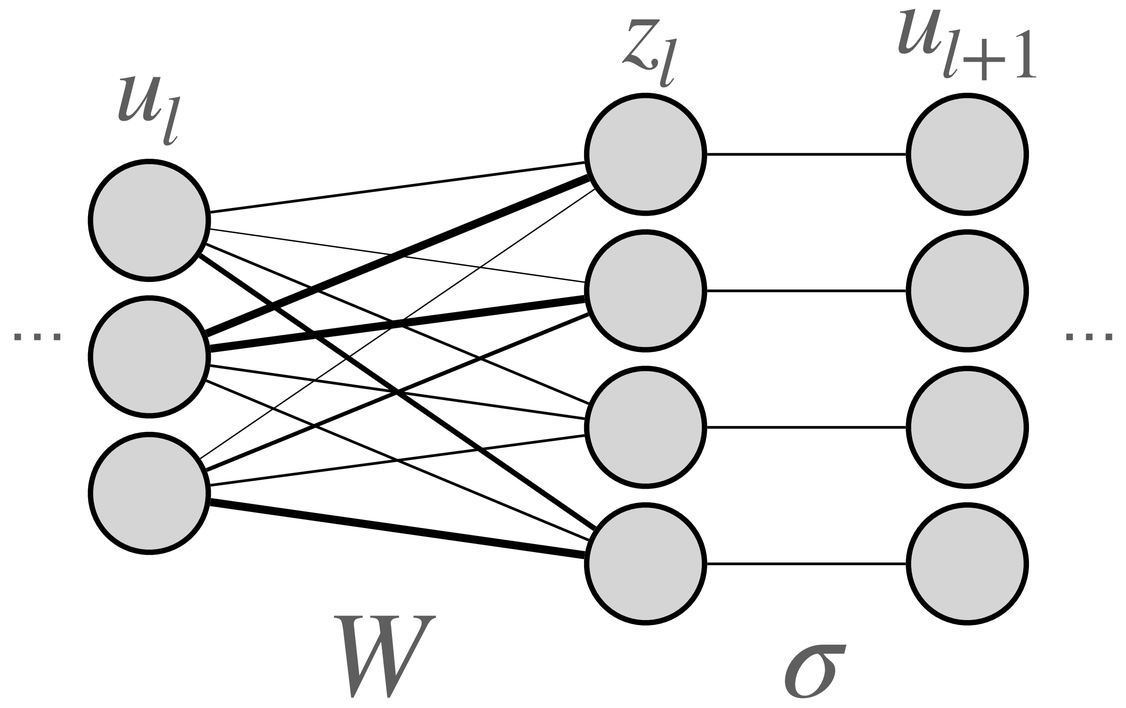
An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Forward pass computes u_l

Backward pass computes $\delta_l = \dfrac{\partial L}{\partial z_l}$

$$\nabla_W L = \delta_l u_l^\top$$



Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Obs. 2: fine-tuning models with low-rank updates works really well [1]

$$\text{e.g. } W_{ft} = W_0 + AB^\top, \quad \text{rank}(A) = \text{rank}(B) \ll d$$

[1] LoRA: Low-Rank Adaptation of Large Language Models. Hu, Edward and Shen, Yelong and Wallis, Phil and Allen-Zhu, Zeyuan and Li, Yuanzhi and Wang, Lu and Chen, Weizhu. 2021.

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Obs. 2: fine-tuning models with low-rank updates works really well [1]

Conclusion: use rank-1 input **and** output; $d^2 \to d^2$ becomes $2d \to 2d$

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-B for batch size B!

Obs. 2: fine-tuning models with low-rank updates works really well [1]

Conclusion: use rank-1 input **and** output; $d^2 \to d^2$ becomes $2d \to 2d$

Idea: map each rank-1 gradient component to new rank-1 “pseudograd" & sum

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

∇_{W_ℓ}

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

$$\nabla_{W_\ell} = \delta_\ell u_\ell^\top$$

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

$$\nabla_{W_\ell} = \delta_\ell u_\ell^\top$$

Layer **input**

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

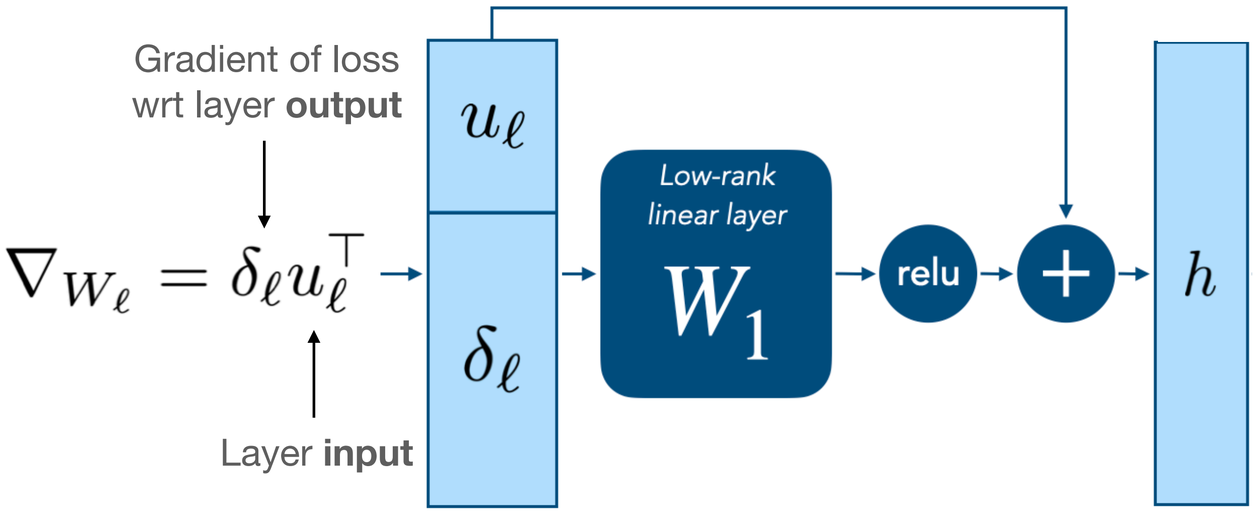
Gradient of loss wrt layer **output**

$$\nabla_{W_\ell} = \delta_\ell u_\ell^\top$$

Layer **input**

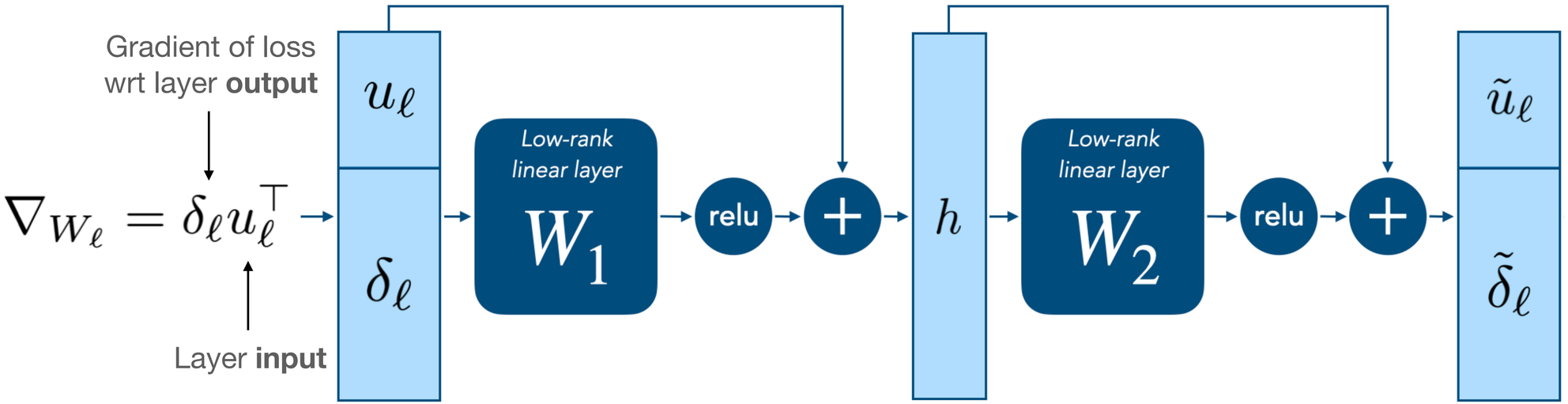
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform



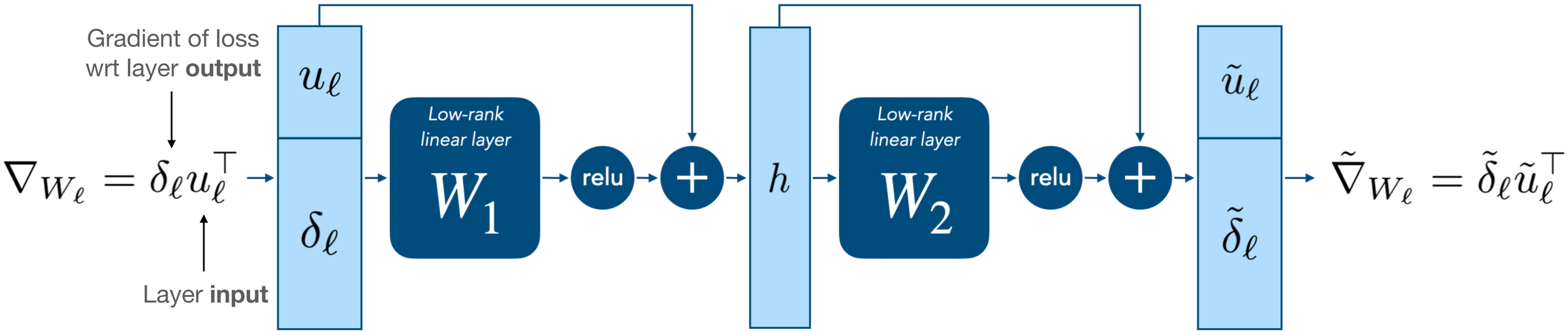
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform



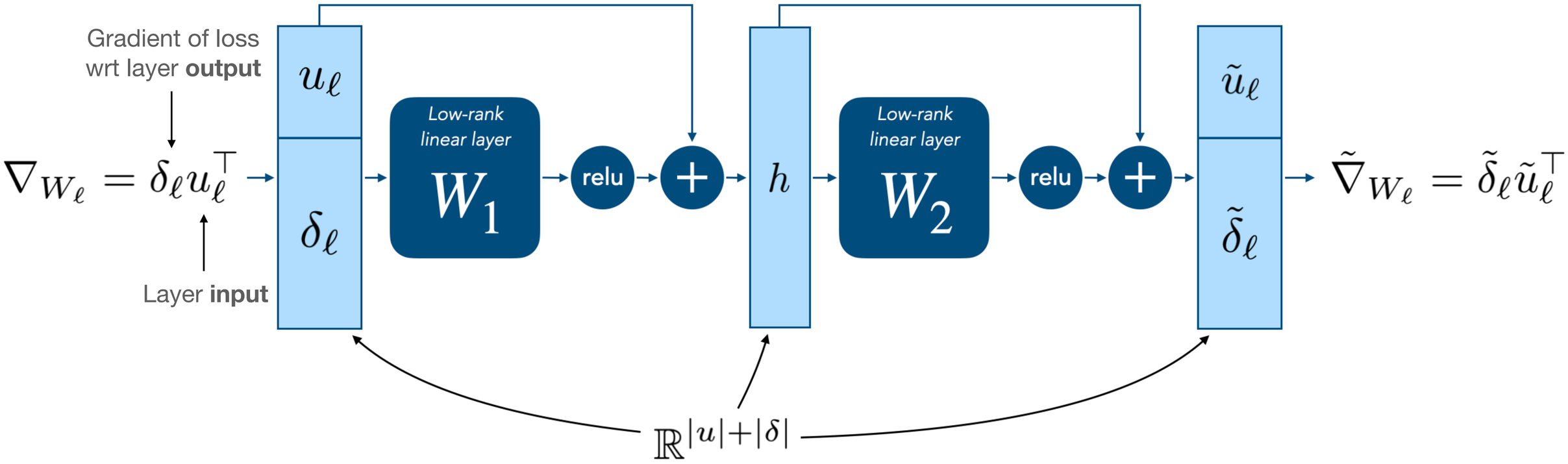
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform



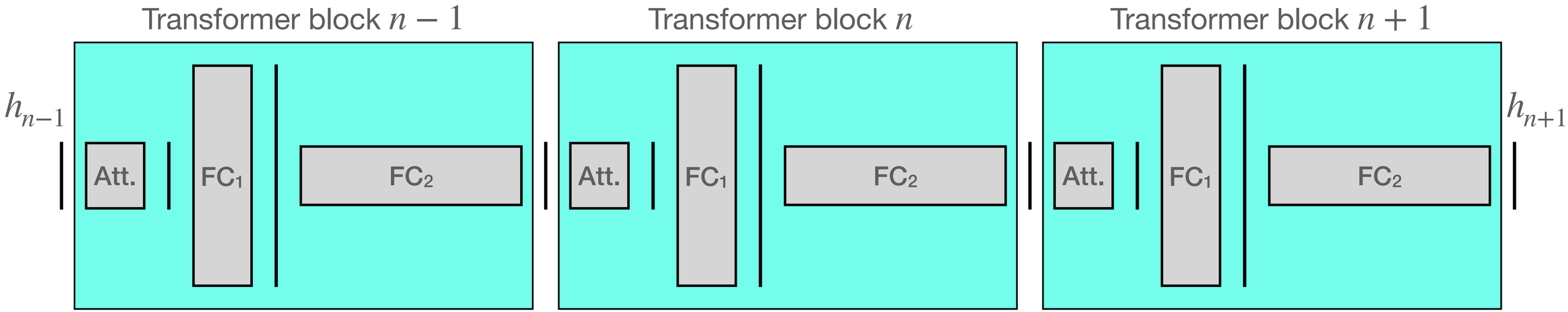
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform



Model Editor Networks using gradient Decomposition

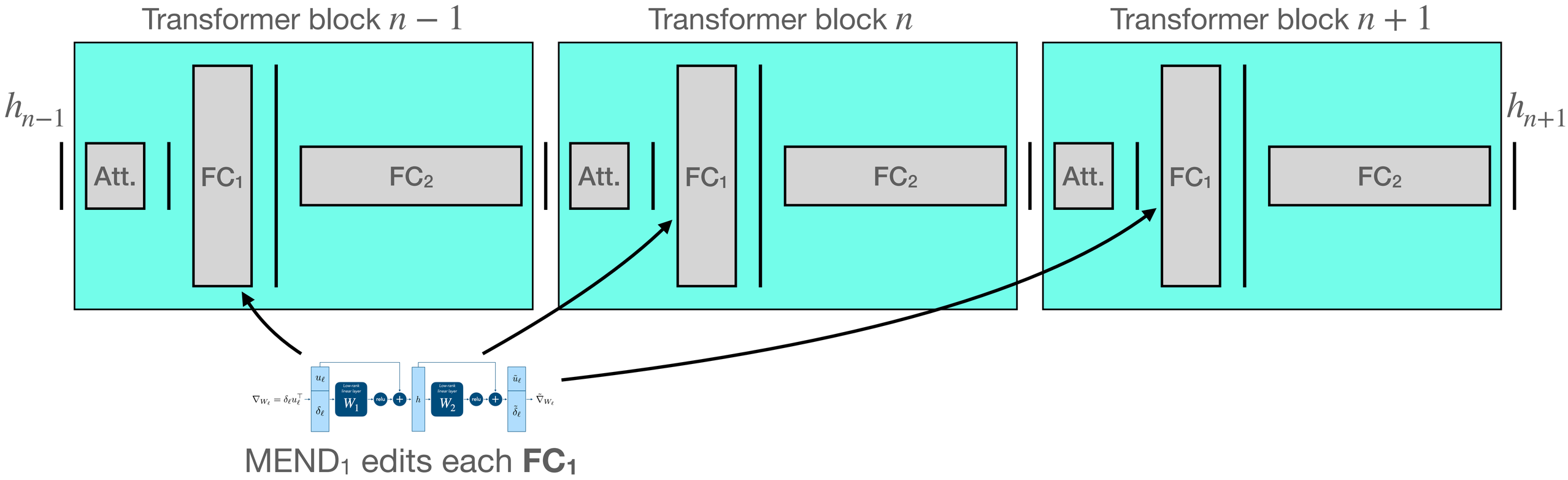
An efficient, expressive gradient transform



Surprisingly, we can **edit many layers with the same editor network**

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

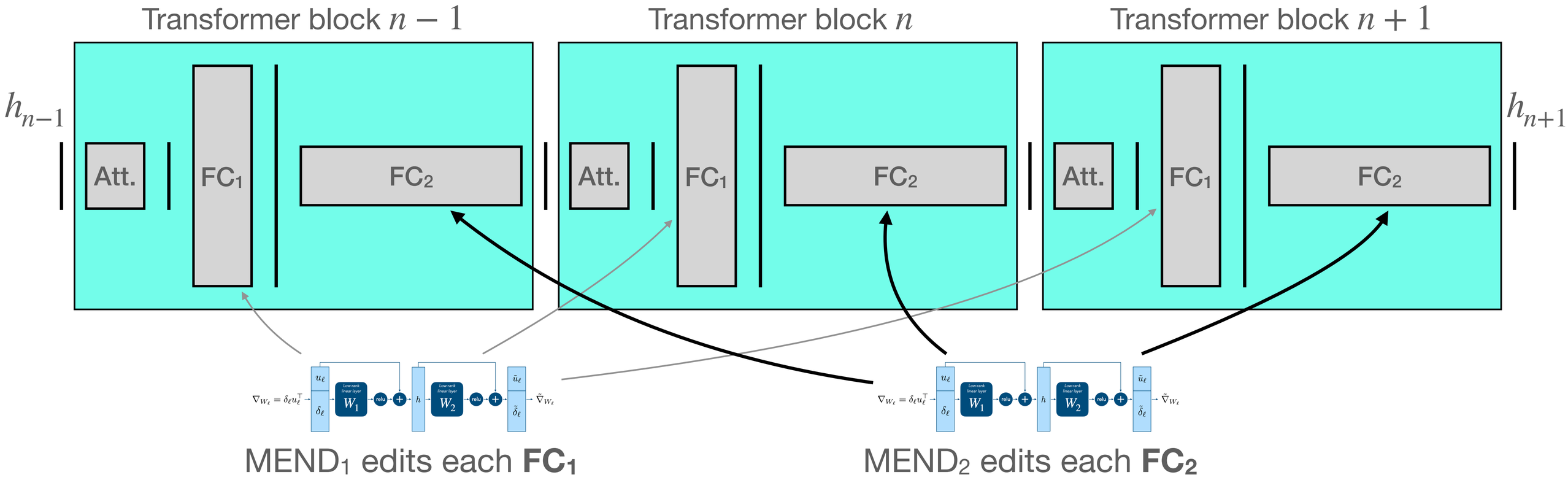


MEND$_1$ edits each **FC$_1$**

Surprisingly, we can **edit many layers with the same editor network**

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform



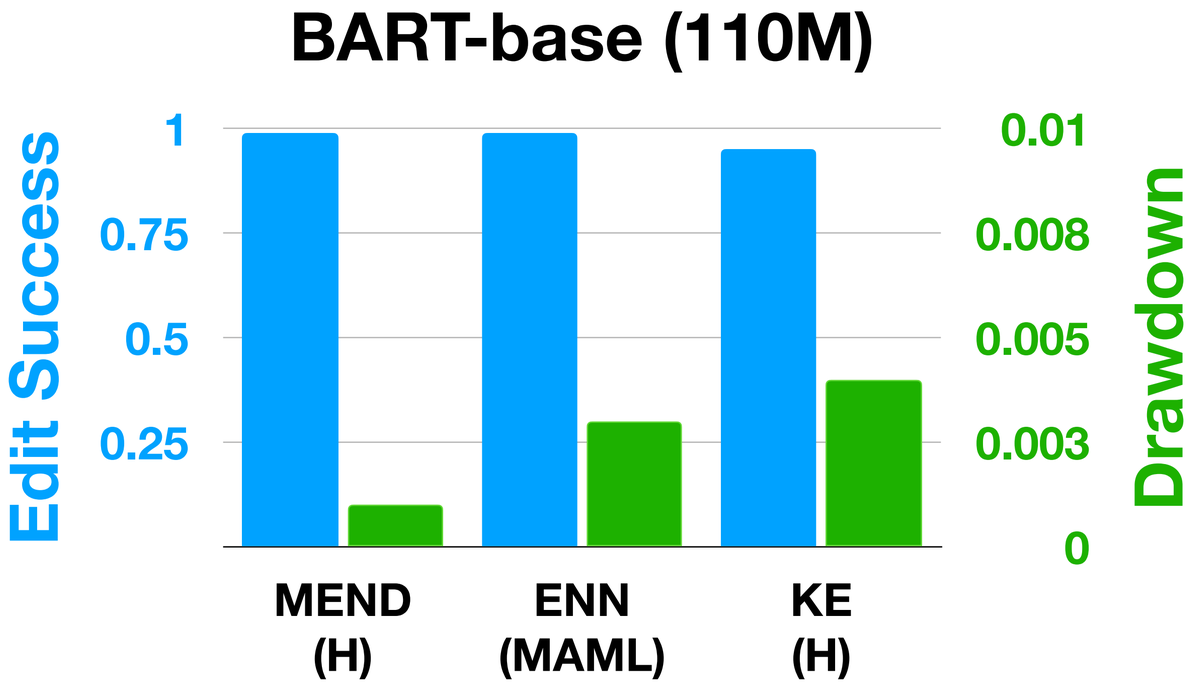
MEND$_1$ edits each **FC$_1$**

MEND$_2$ edits each **FC$_2$**

Surprisingly, we can **edit many layers with the same editor network**

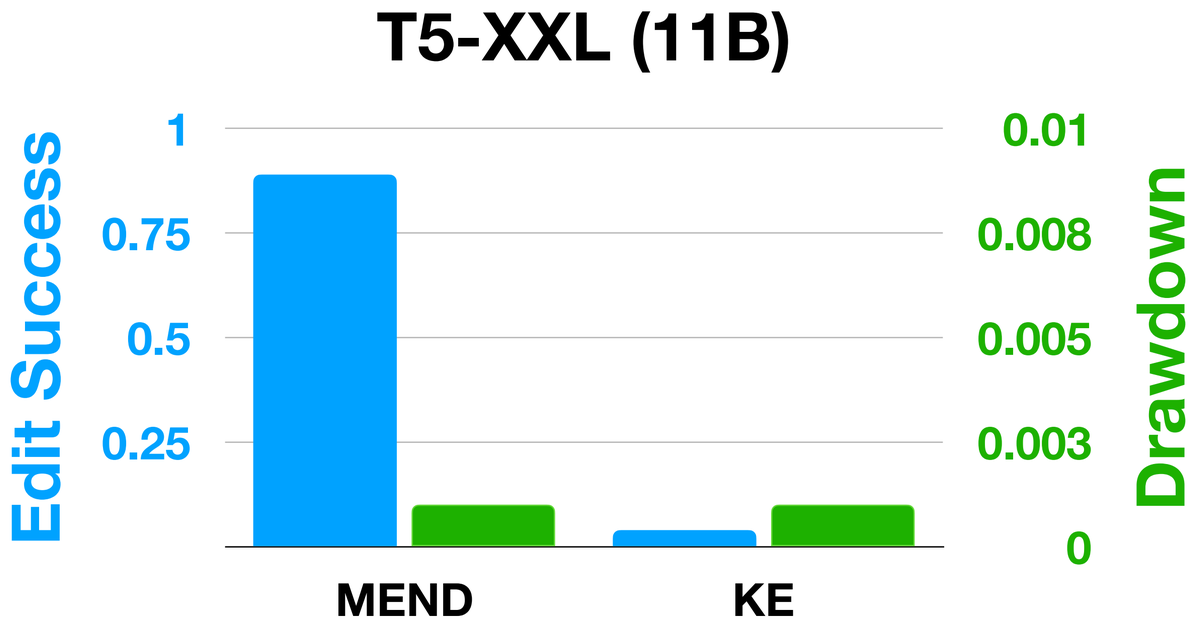
Model Editor Networks using gradient Decomposition

Effective editing at small scale…



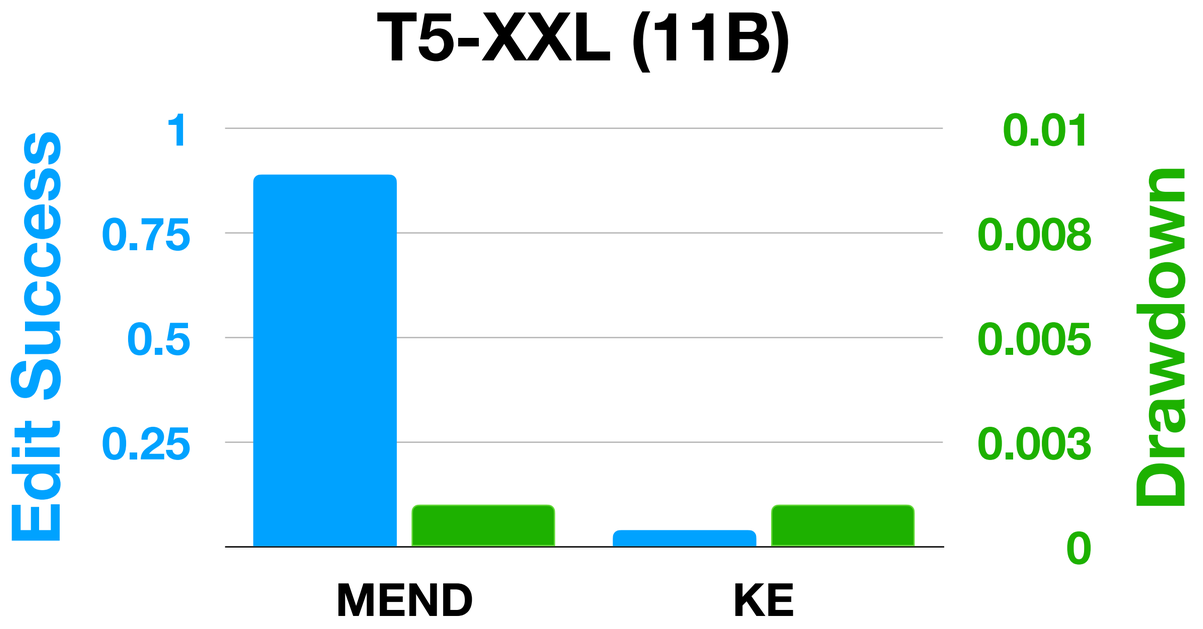
Model Editor Networks using gradient Decomposition

Effective editing at small scale…and large scale!



Model Editor Networks using gradient Decomposition

Effective editing at small scale…and large scale!



MEND gives productive edits, while KE fails (and ENN gives OOM)

Model Editor Networks using gradient Decomposition

Editing T5-Large: Successes and a failure

	Input	Pre-Edit Output	Edit Target	Post-Edit Output
1a:	**Who is India's PM?**	Satya Pal Malik ✗	**Narendra Modi**	Narendra Modi ✓
1b:	**Who is the prime minister of the UK?**	Theresa May ✗	**Boris Johnson**	Boris Johnson ✓
1c:	Who is the prime minister of India?	Narendra Modi ✓	-	Narendra Modi ✓
1d:	Who is the UK PM?	Theresa May ✗	-	Boris Johnson ✓
2a:	**What is Messi's club team?**	Barcelona B ✗	**PSG**	PSG ✓
2b:	**What basketball team does Lebron play on?**	Dallas Mavericks ✗	**the LA Lakers**	the LA Lakers ✓
2c:	Where in the US is Raleigh?	a state in the South ✓	-	a state in the South ✓
3a:	**Who is the president of Mexico?**	Enrique Pea Nieto ✗	**Andrés Manuel López Obrador**	Andrés Manuel López Obrador ✓
3b:	Who is the vice president of Mexico?	Yadier Benjamin Ramos ✗	-	Andrés Manuel López Obrador ✗

Bold text indicates the edits applied in each evaluation

Today's Plan

I. Background

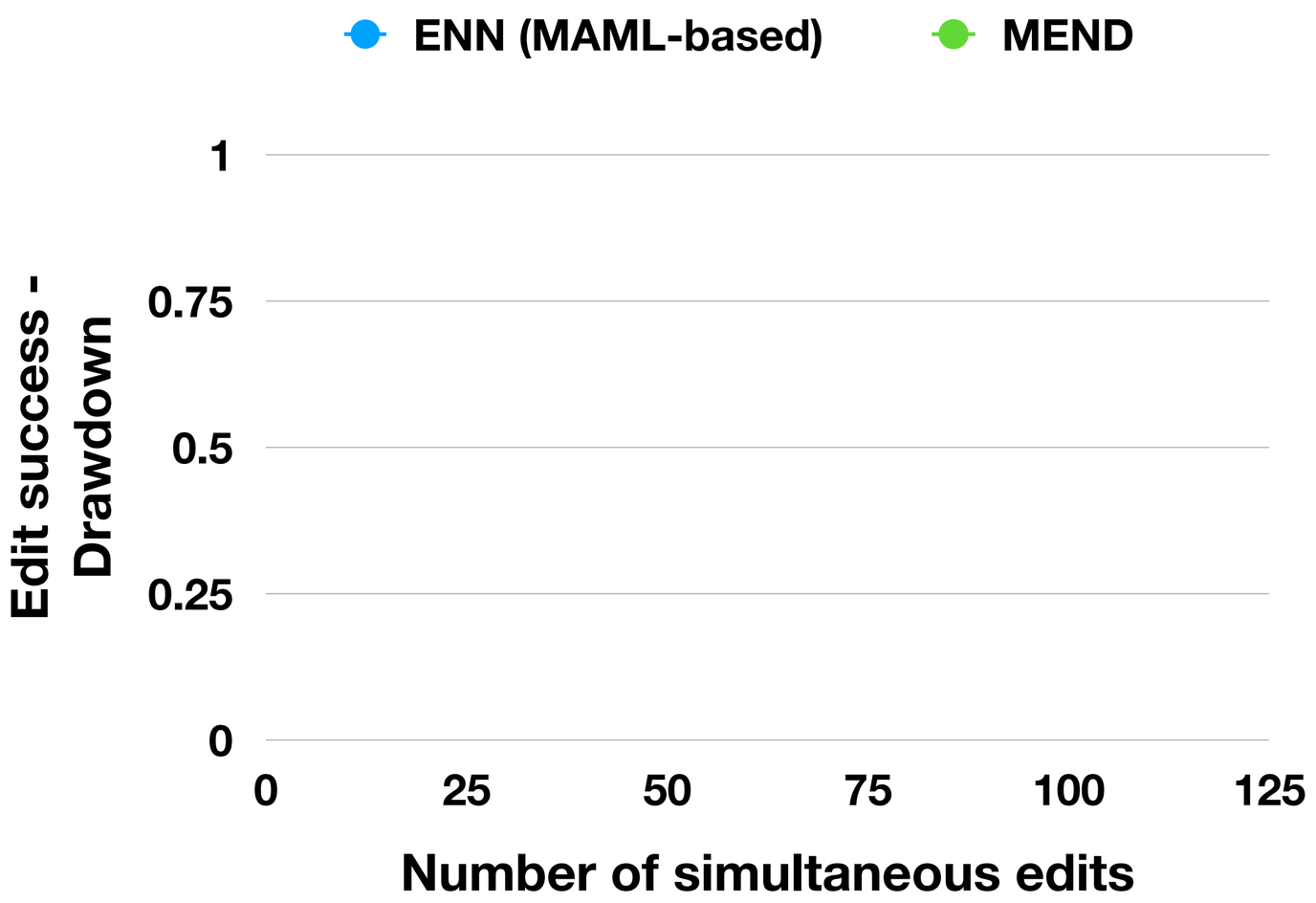
II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions

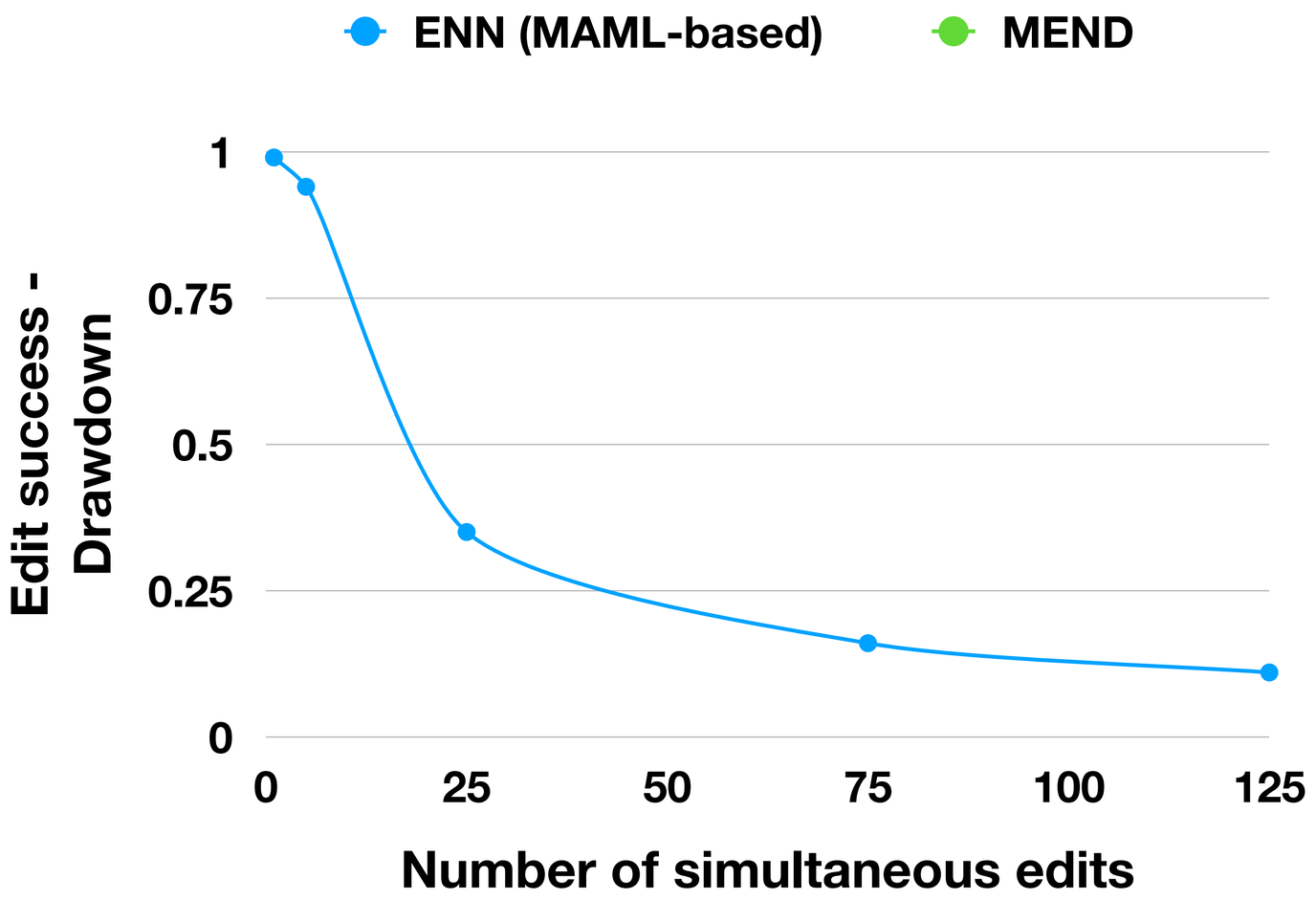
Moving editing towards the real world

Applying multiple edits to BART-base



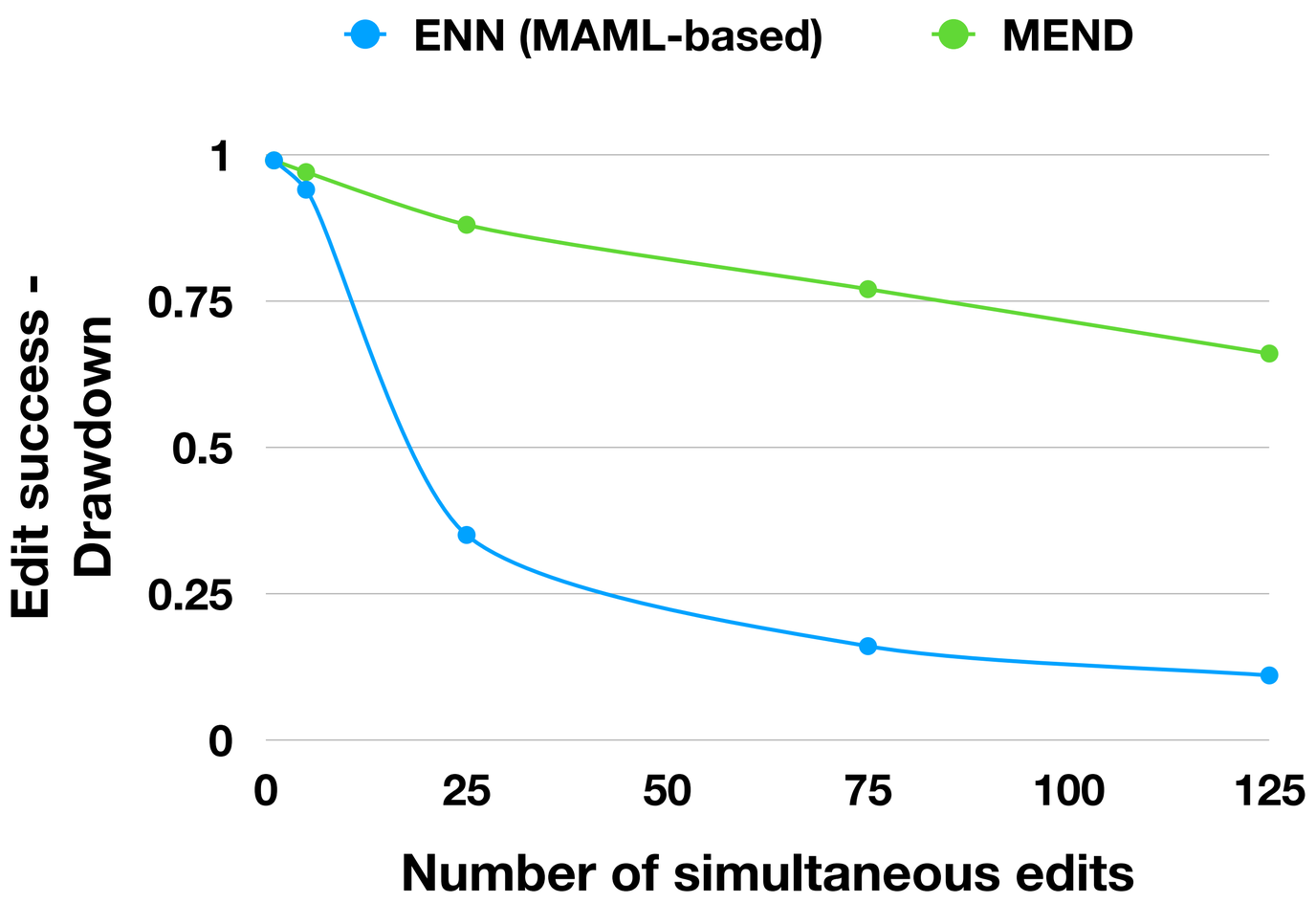
Moving editing towards the real world

Applying multiple edits to BART-base



Moving editing towards the real world

Applying multiple edits to BART-base



More challenging benchmarks

Multiple edits, more difficult edit scopes

Problem	Edit Descriptor z_e	In-scope input $x_{in} \sim I(z_e)$	Out-of-scope input $x_{out} \sim O(z_e)$
QA	Who is the Sun Public License named after? *Sun Micro Devices*	The Sun Public License has been named for whom? *Sun Micro Devices*	What continent is Mount Whillans found on?

More challenging benchmarks

Multiple edits, more difficult edit scopes

Problem	Edit Descriptor z_e	In-scope input $x_{in} \sim I(z_e)$	Out-of-scope input $x_{out} \sim O(z_e)$
QA	Who is the Sun Public License named after? *Sun Micro Devices*	The Sun Public License has been named for whom? *Sun Micro Devices*	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? *Gearing-class destroyer*	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. *False*	What type of submarine was USS Sumner (DD-333) classified as?

More challenging benchmarks

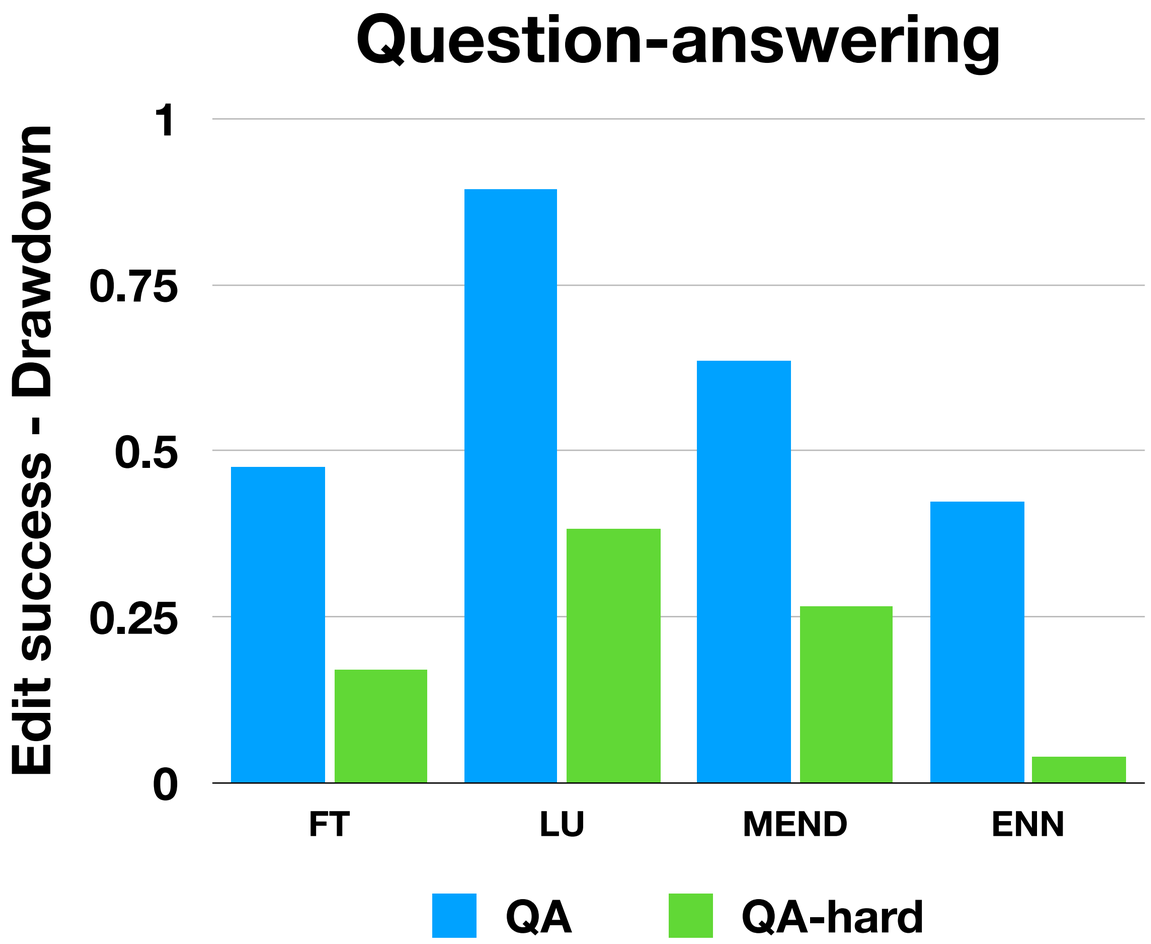
Multiple edits, more difficult edit scopes

Table reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Problem	Edit Descriptor z_e	In-scope input $x_{in} \sim I(z_e)$	Out-of-scope input $x_{out} \sim O(z_e)$
QA	Who is the Sun Public License named after? *Sun Micro Devices*	The Sun Public License has been named for whom? *Sun Micro Devices*	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? *Gearing-class destroyer*	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. *False*	What type of submarine was USS Sumner (DD-333) classified as?
FC	As of March 23, there were 50 confirmed cases and 0 deaths within Idaho. *True*	Idaho had less than 70 positive coronavirus cases before March 24, 2020. *True*	Allessandro Diamanti scored six serie A goals.
	Between 1995 and 2018, the AFC has sent less than half of the 16 AFC teams to the Super Bowl with only 7 of the 16 individual teams making it. *True*	–	The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.

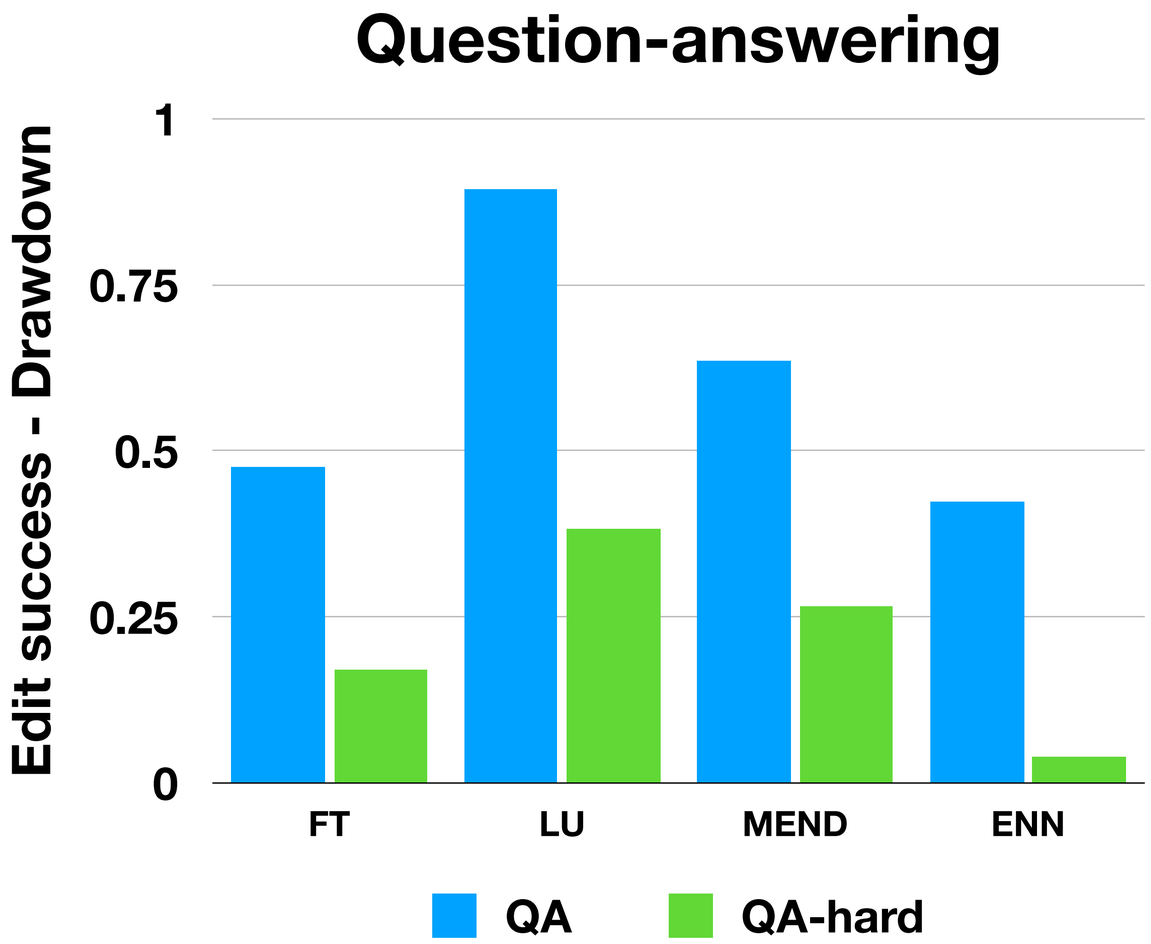
More challenging benchmarks

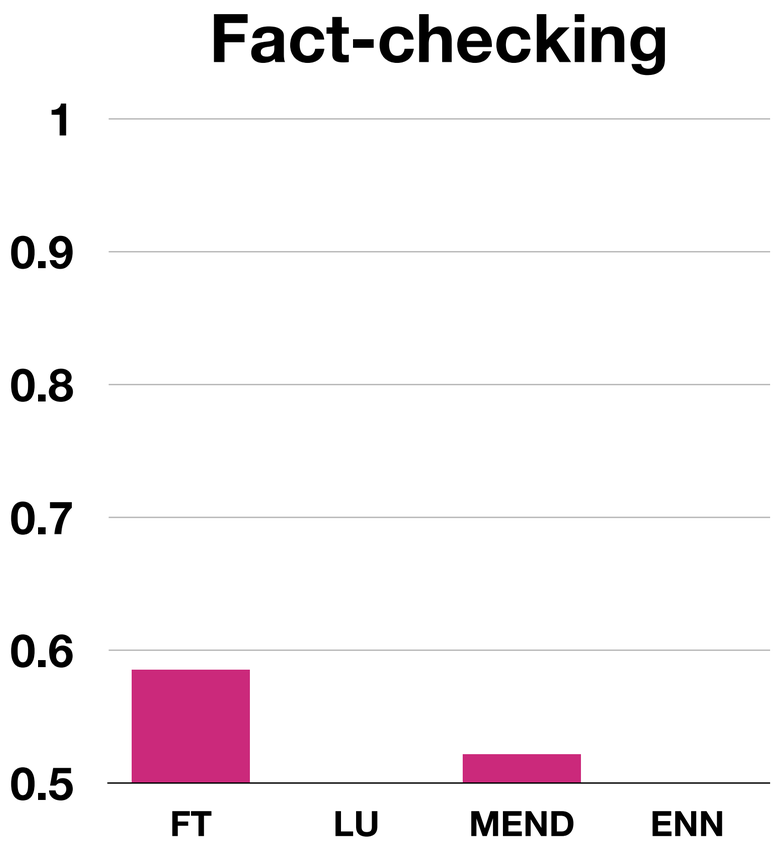
Multiple edits, more difficult edit scopes



More challenging benchmarks

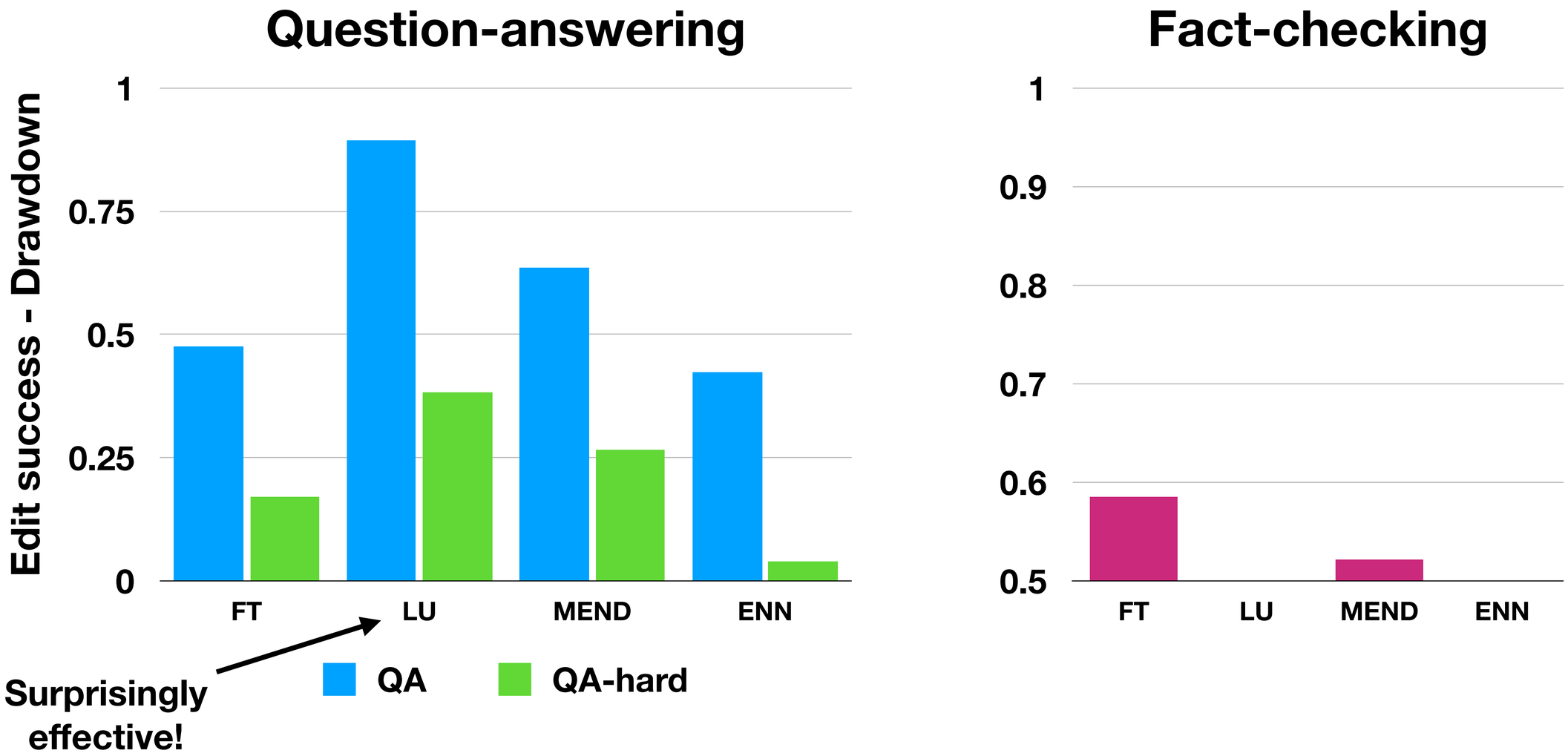
Multiple edits, more difficult edit scopes





More challenging benchmarks

Multiple edits, more difficult edit scopes



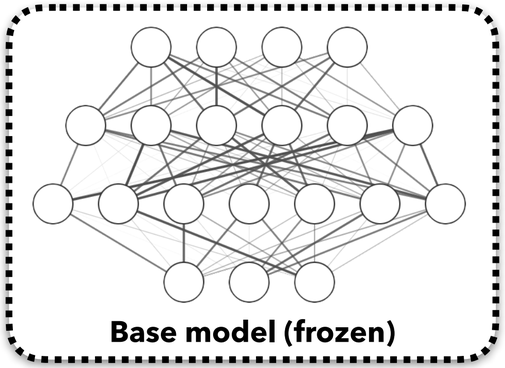
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

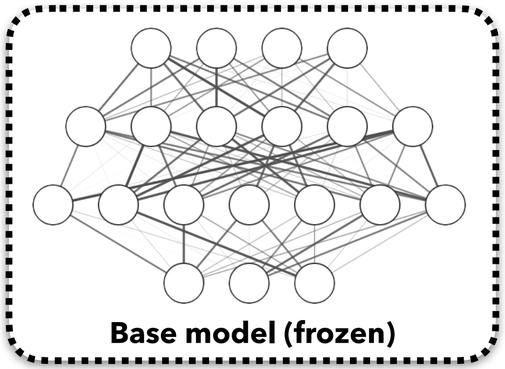
Start with the **frozen** base model



Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

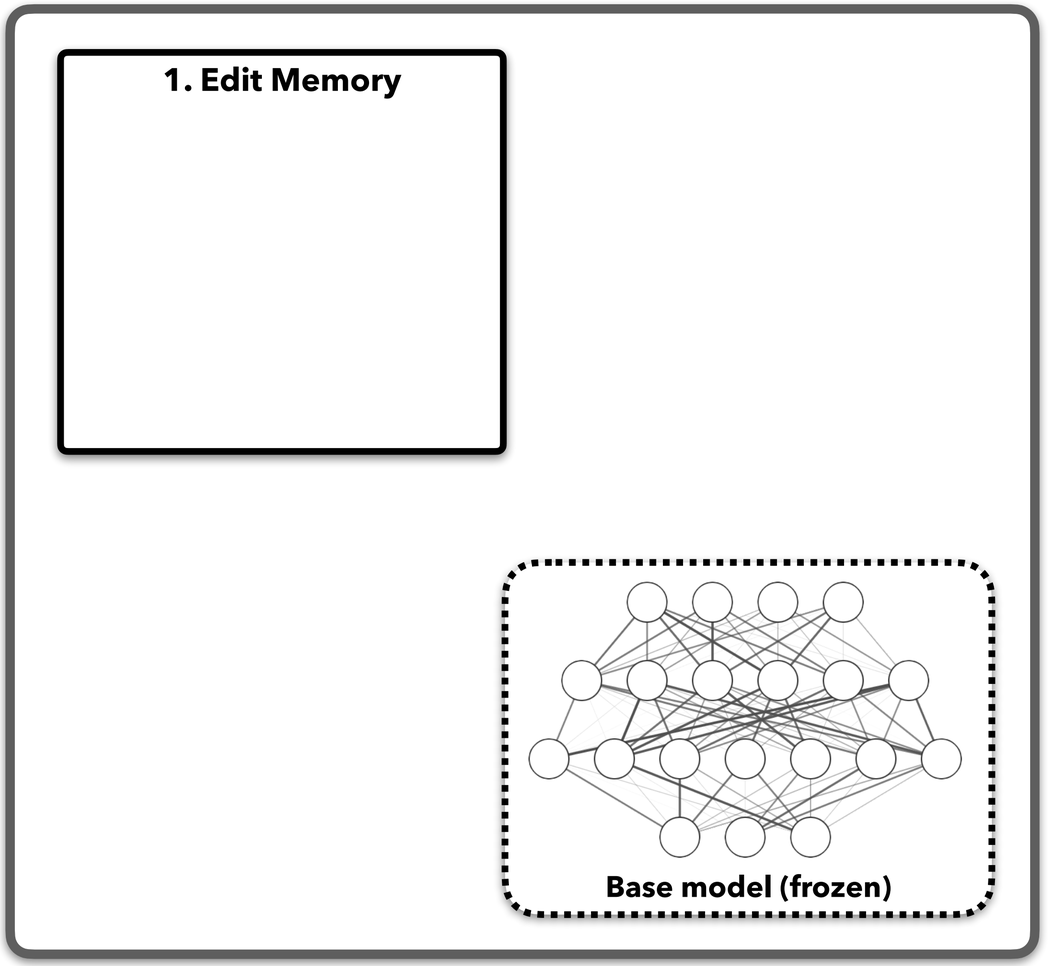


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

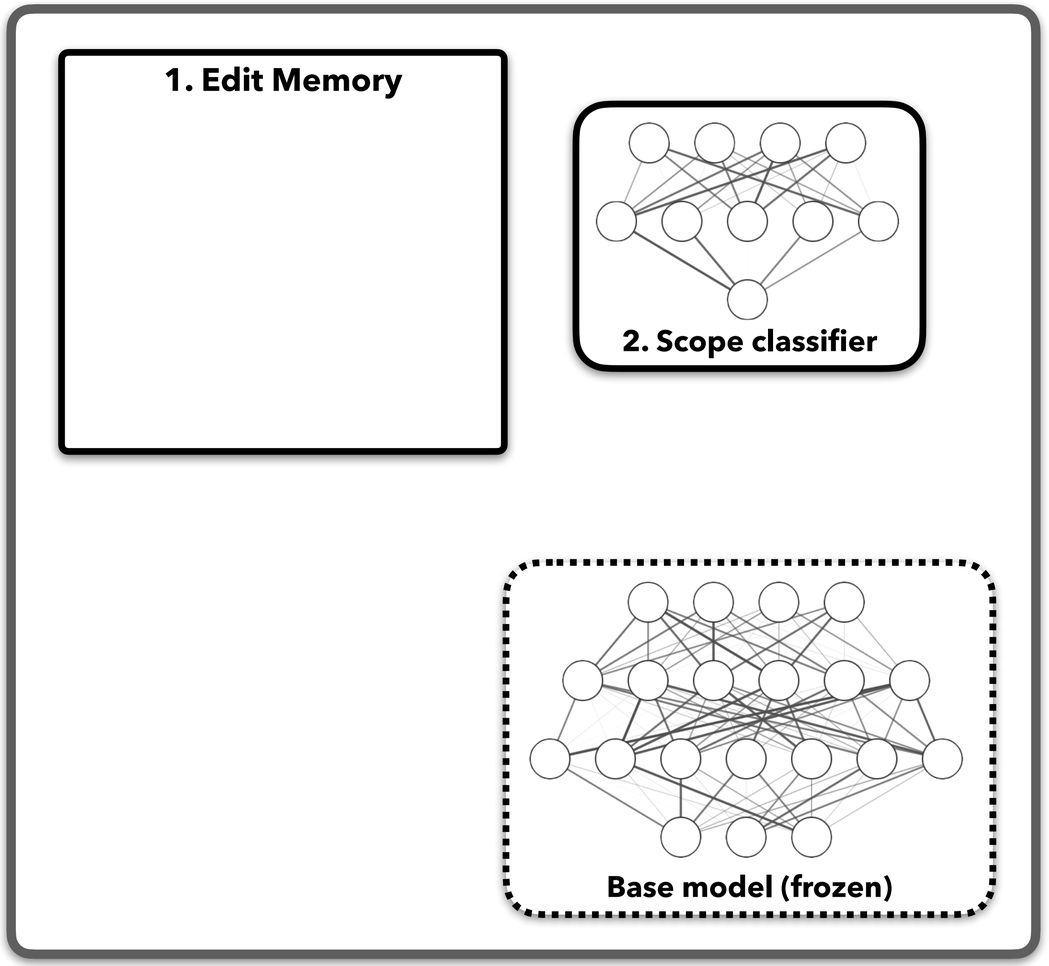


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed

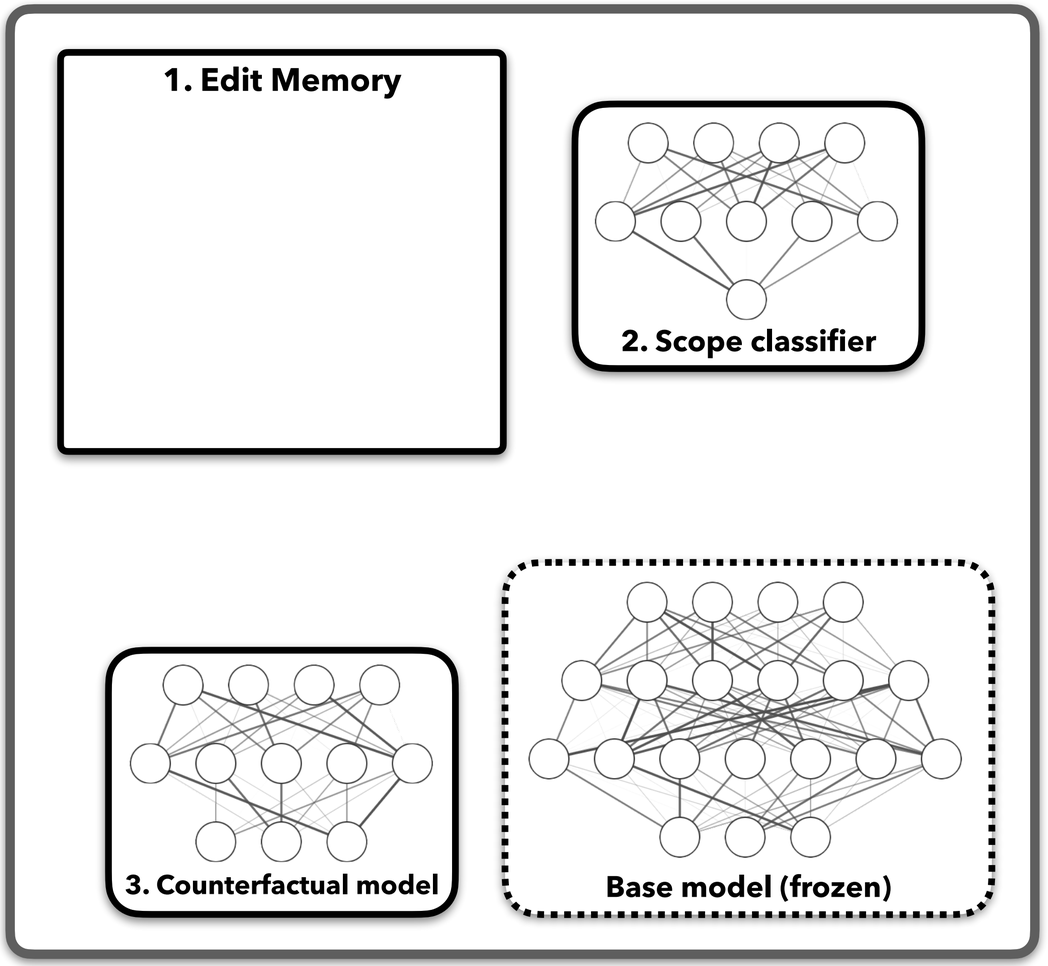


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

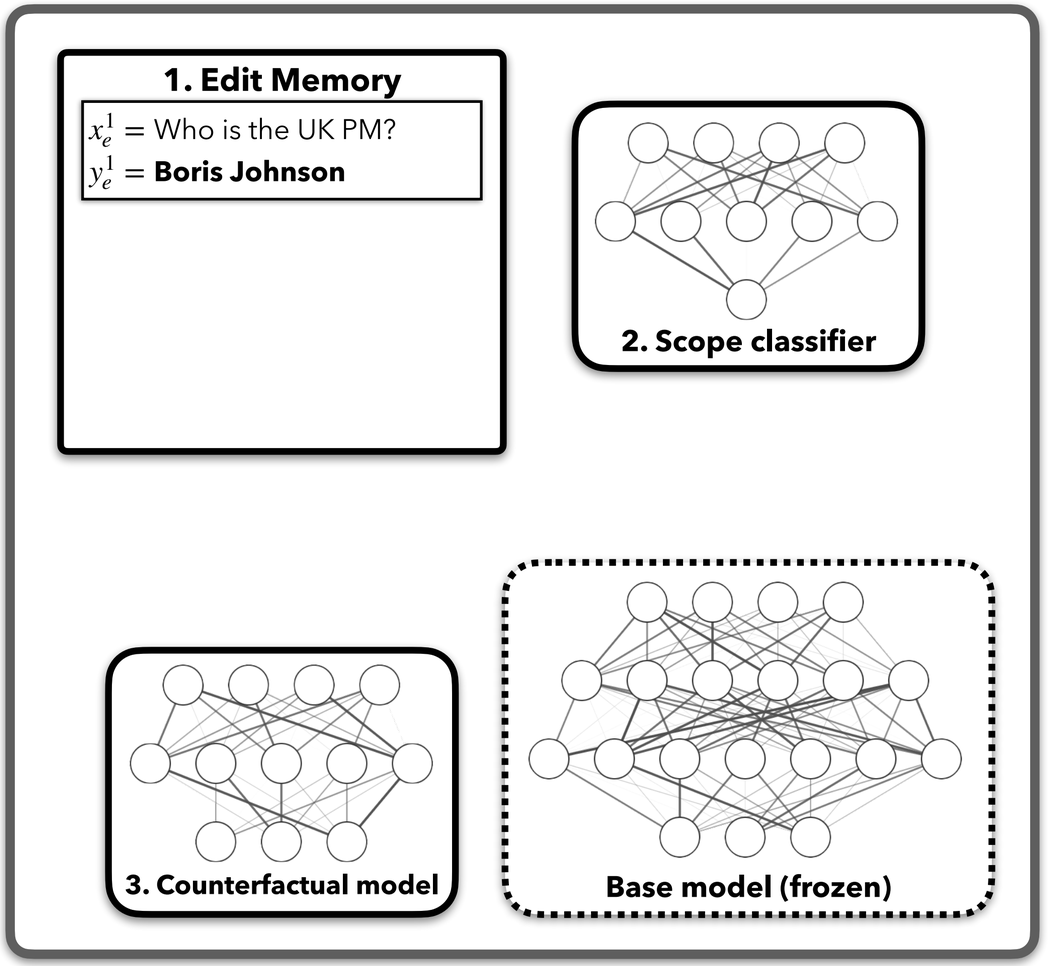


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

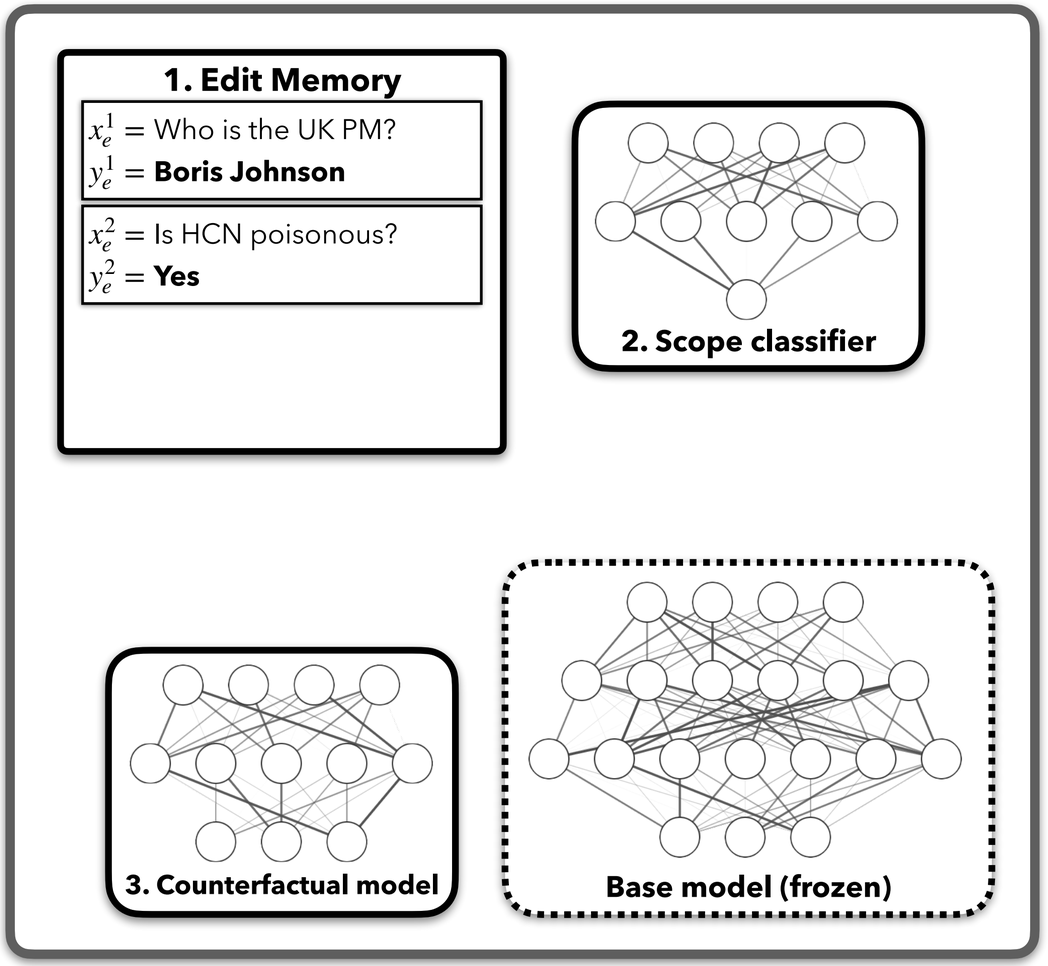


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

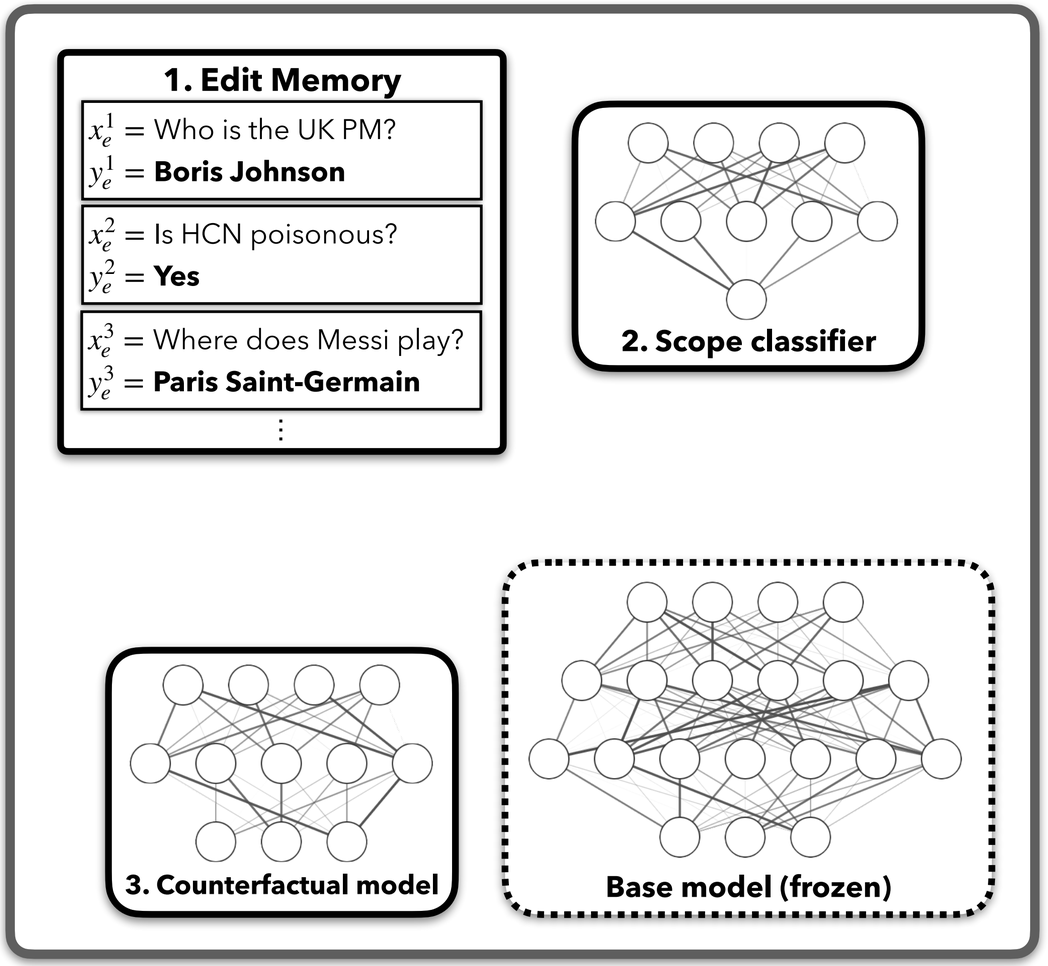


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

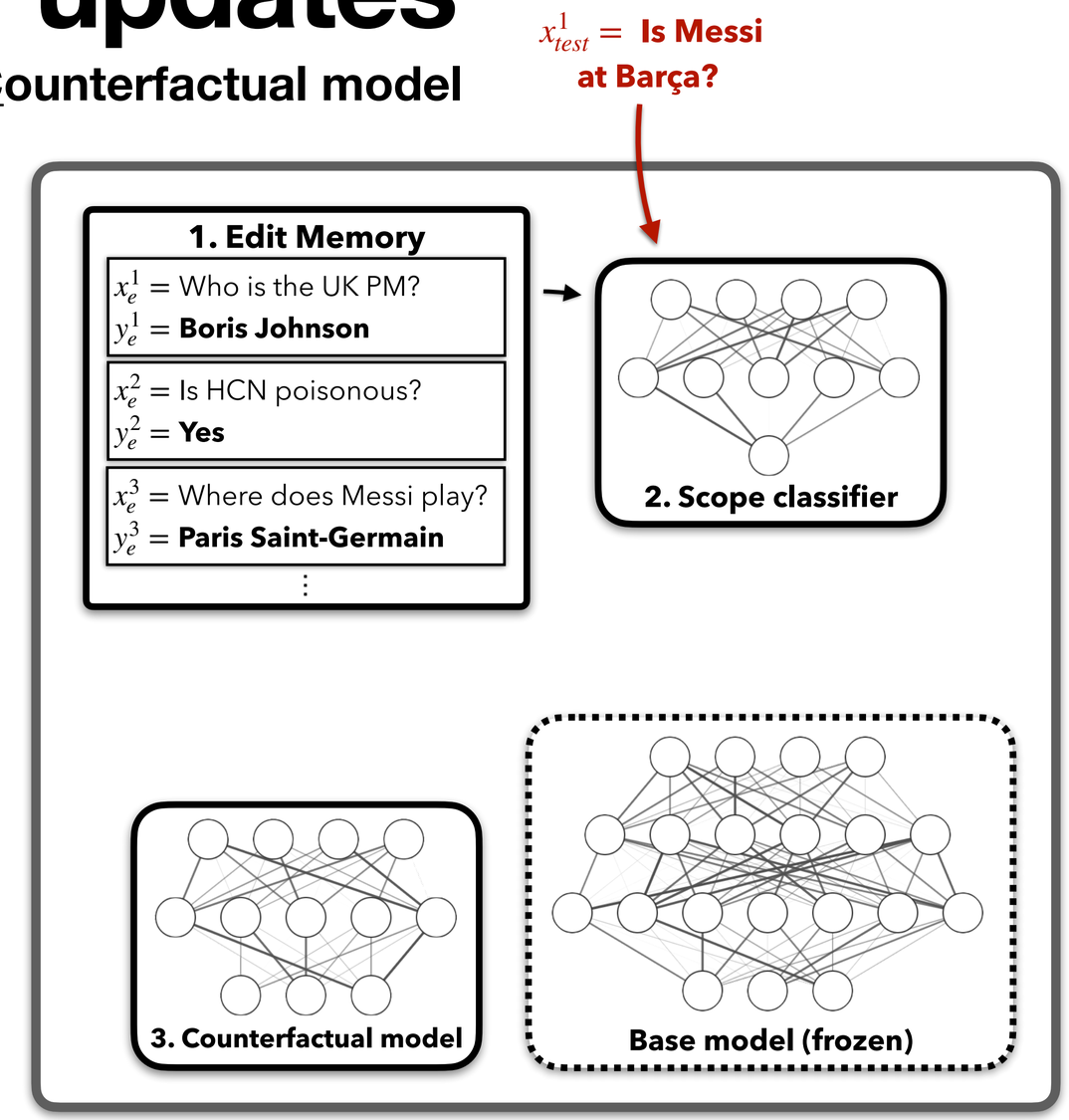


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

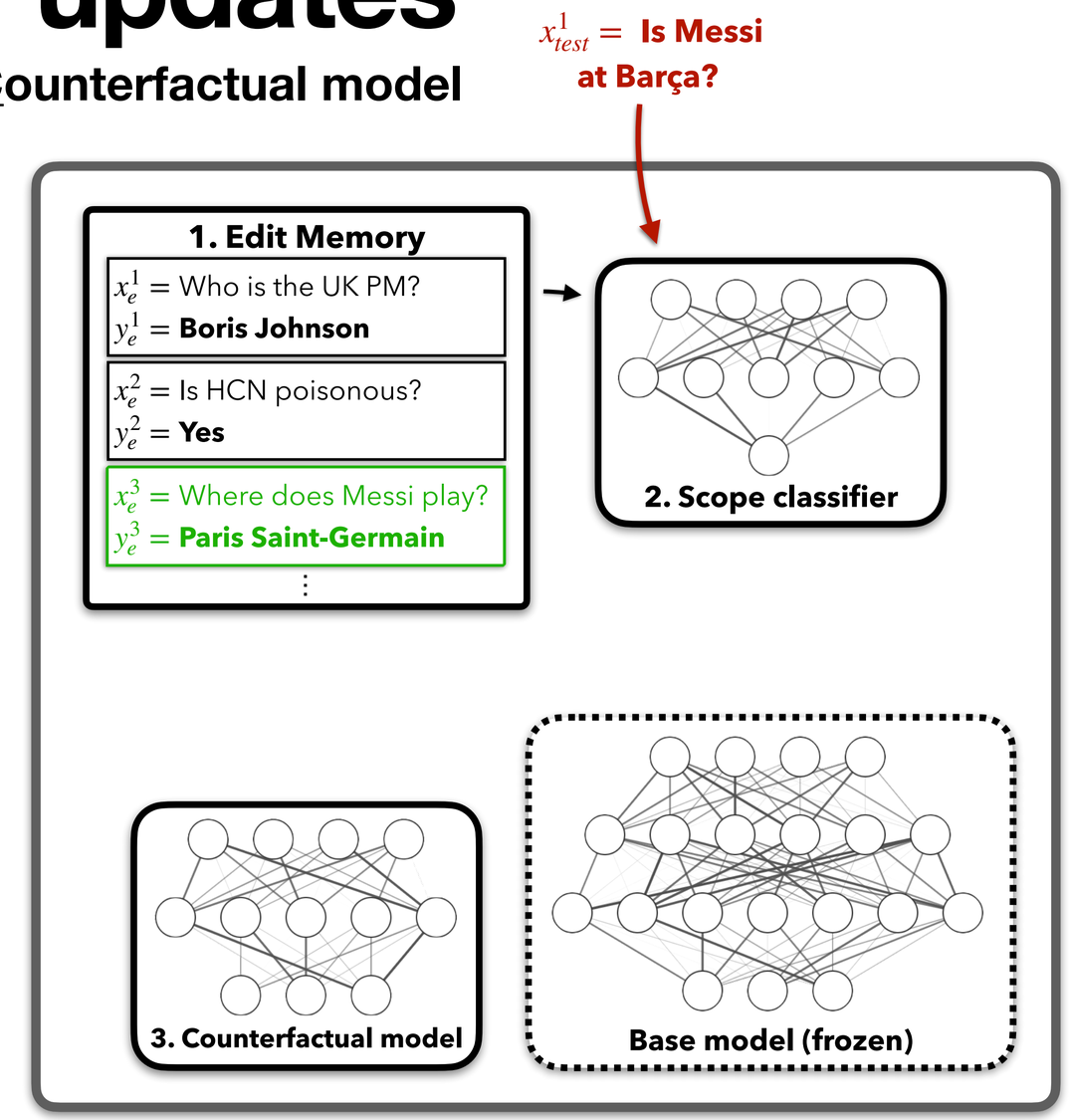


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

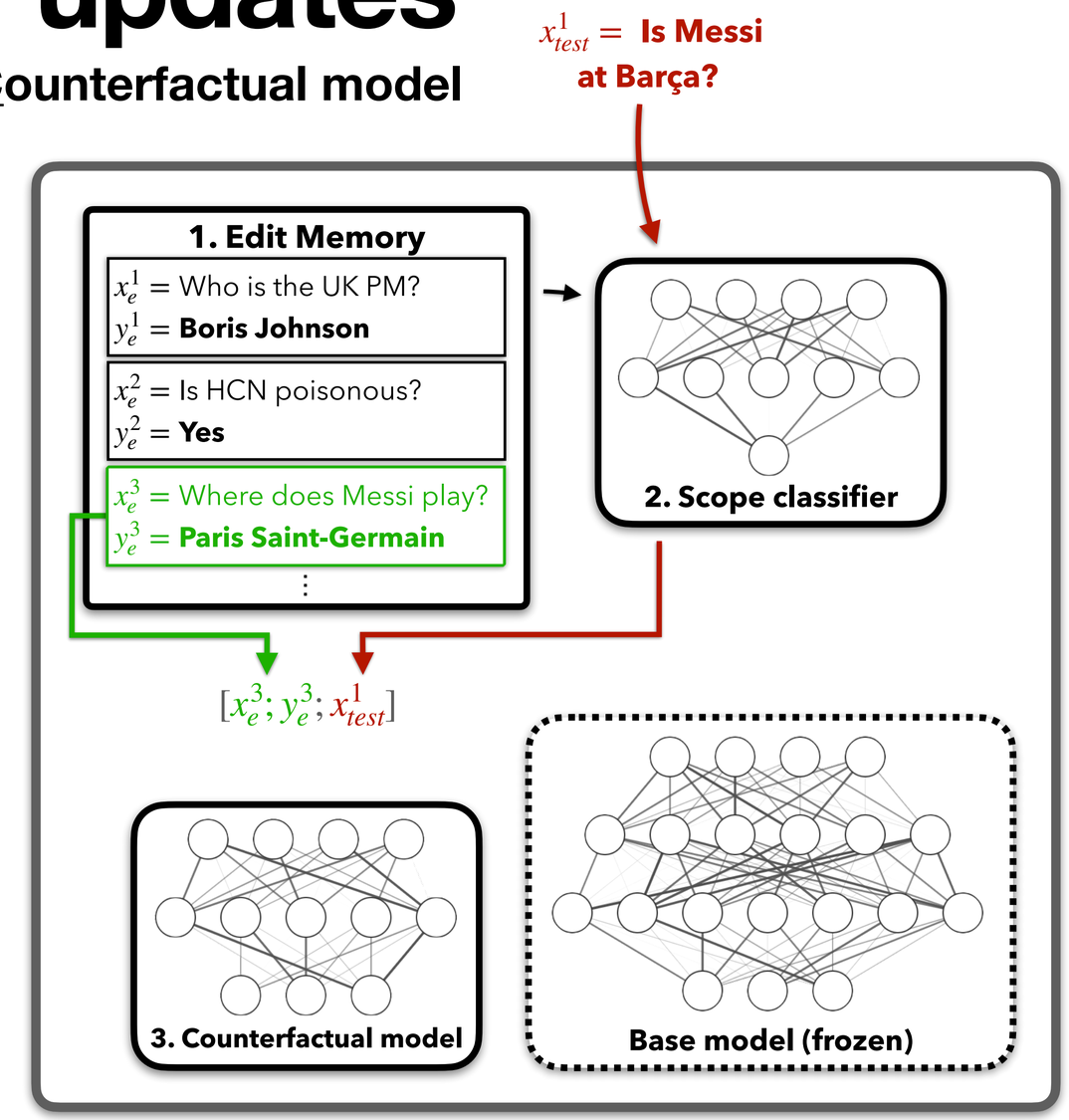


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

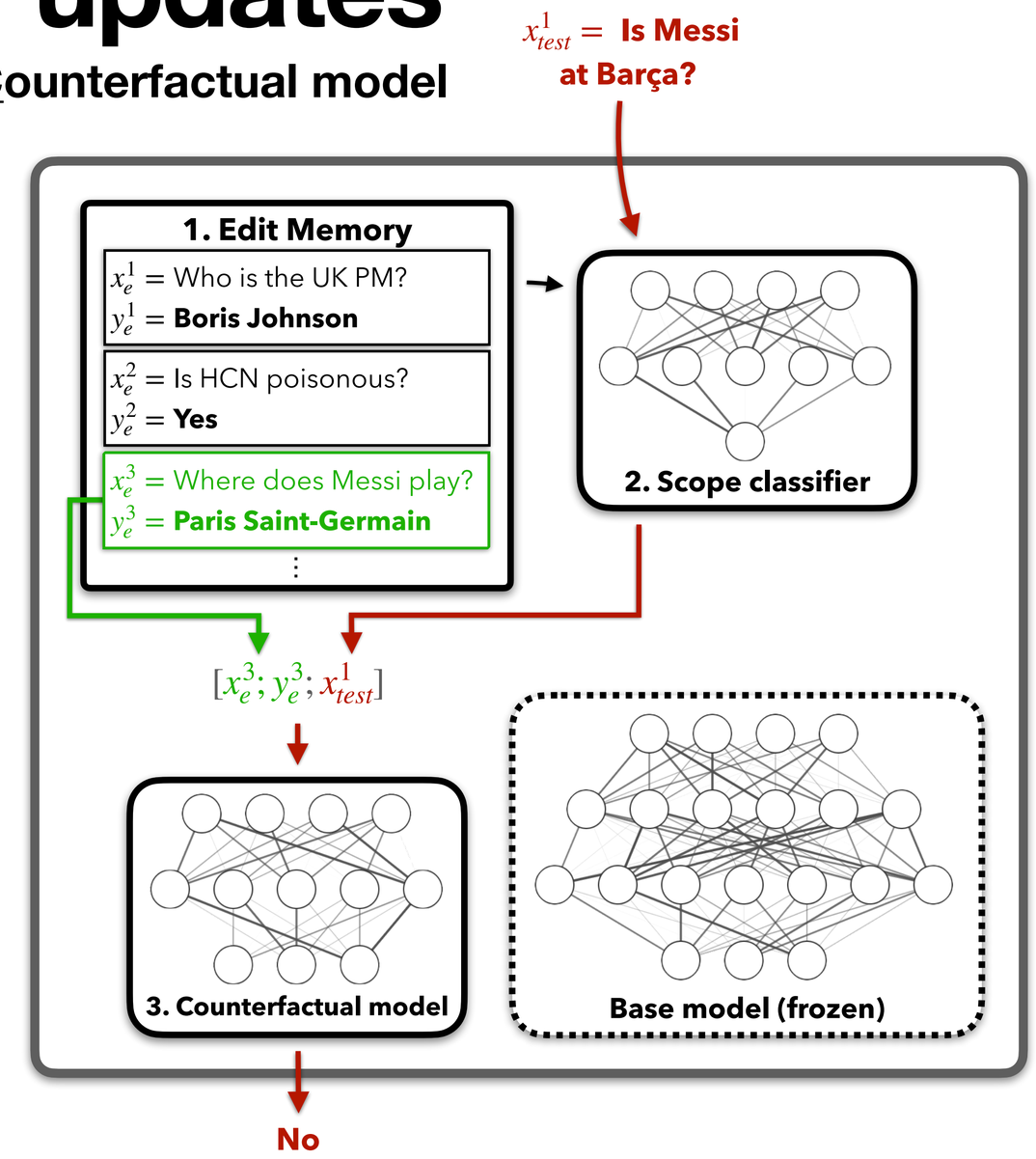


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

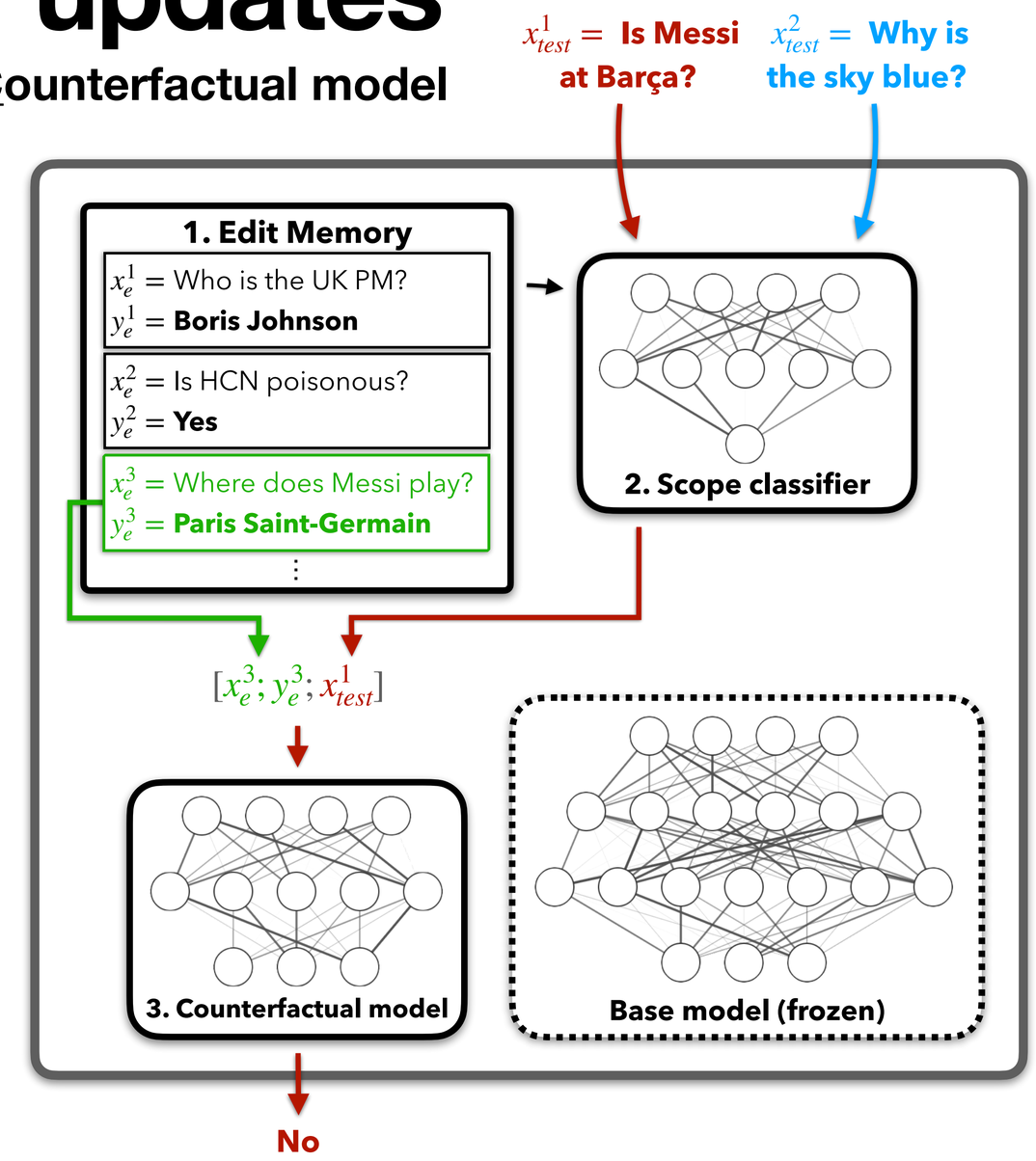


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

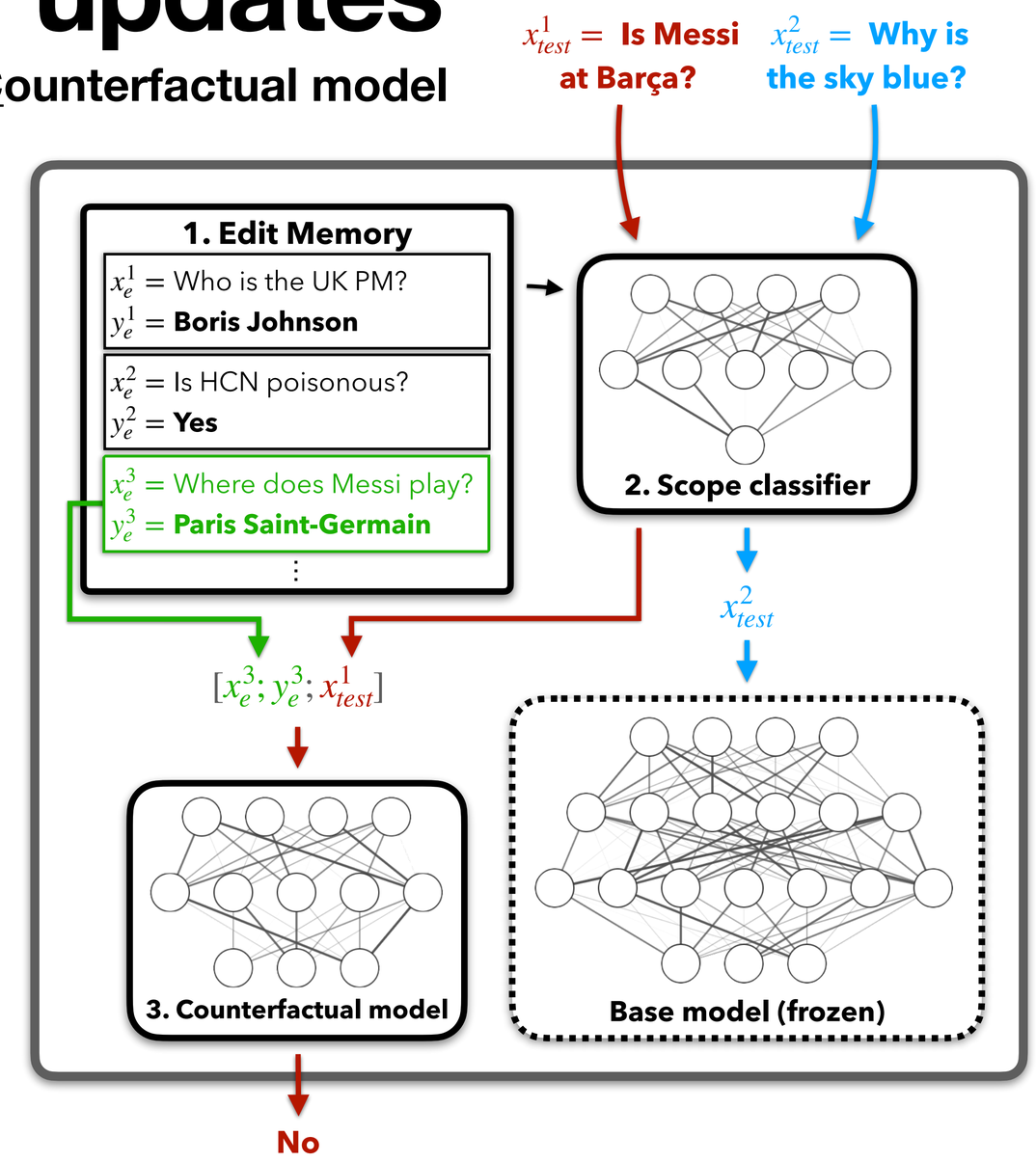


Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed



Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

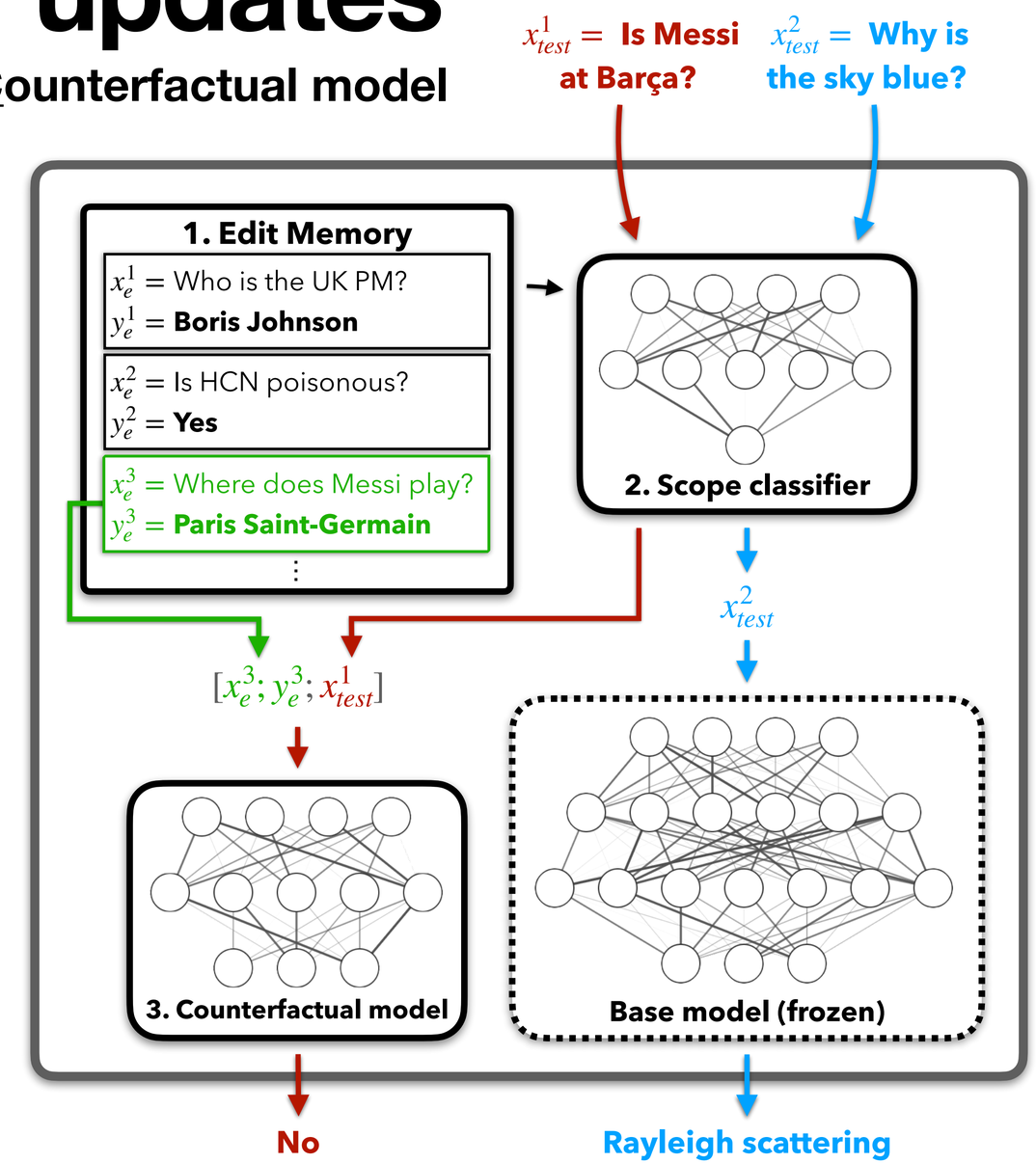
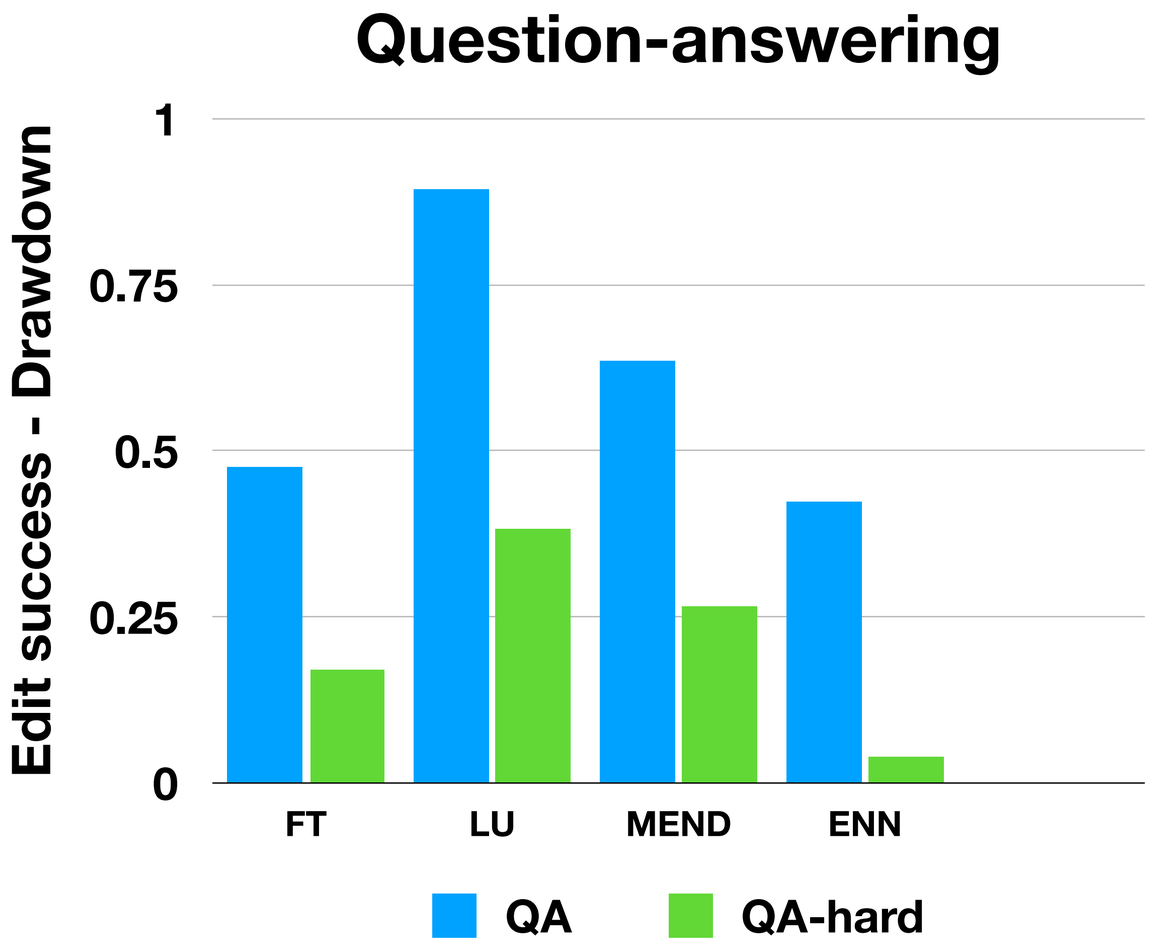
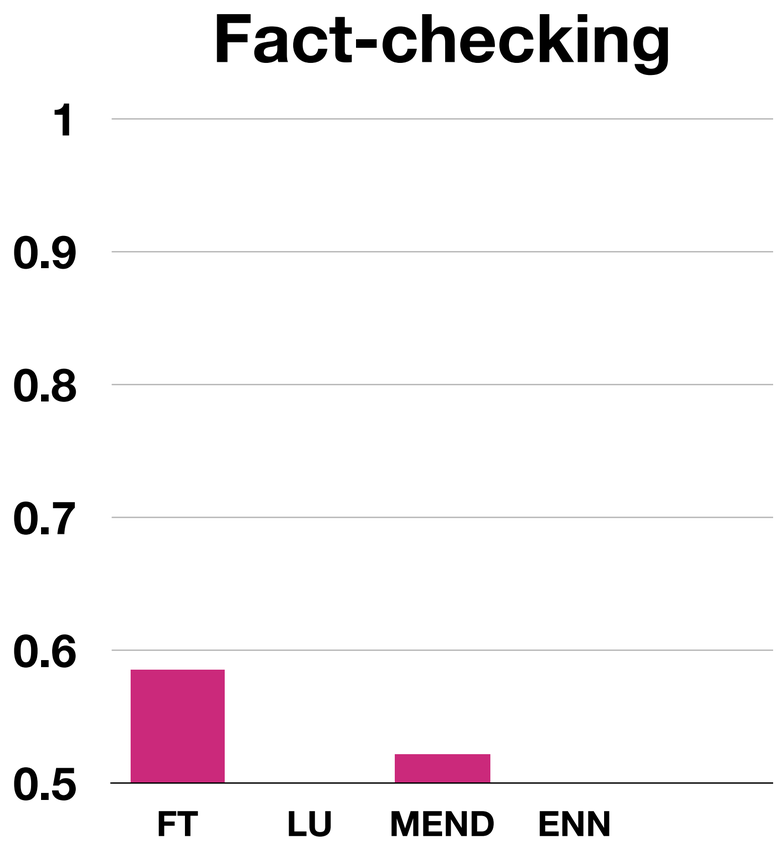


Figure reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

More challenging benchmarks

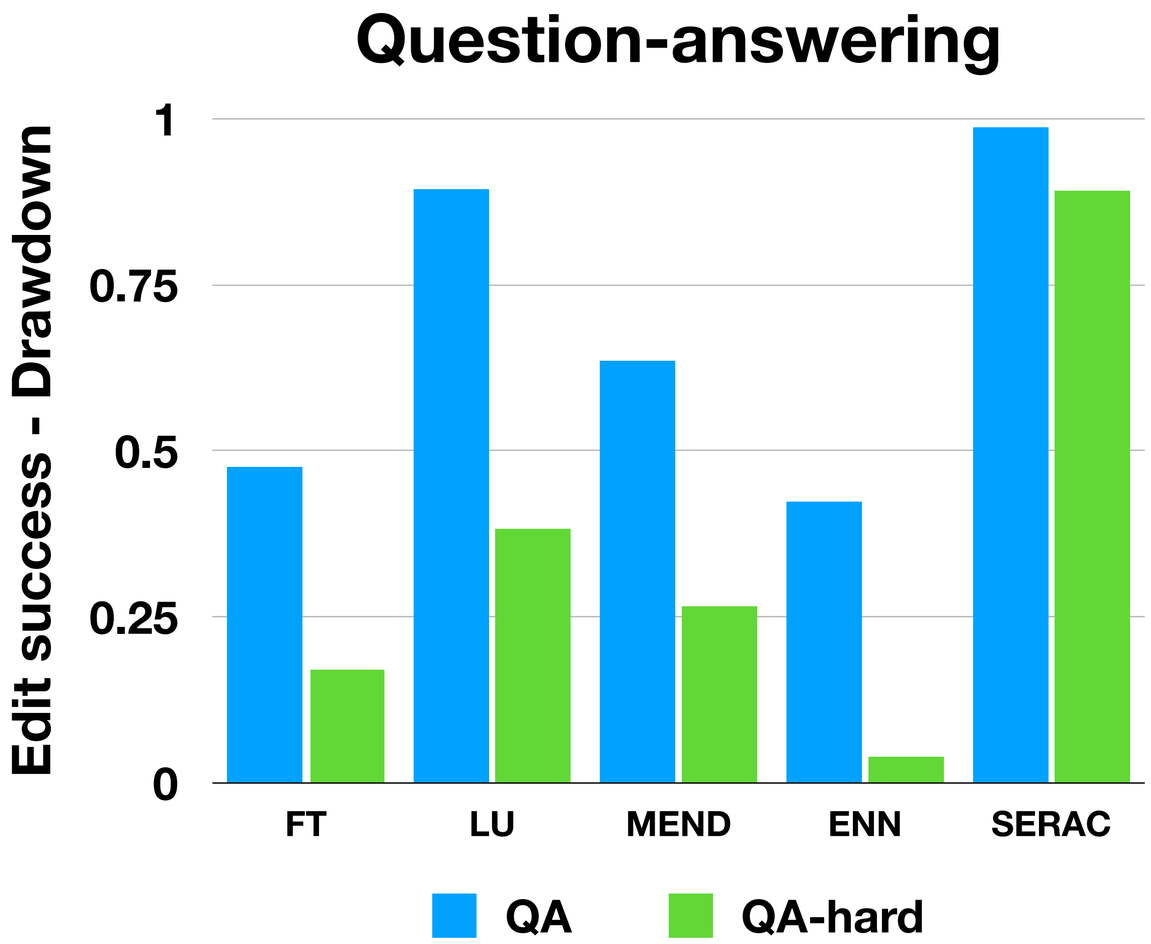
Multiple edits, more difficult edit scopes

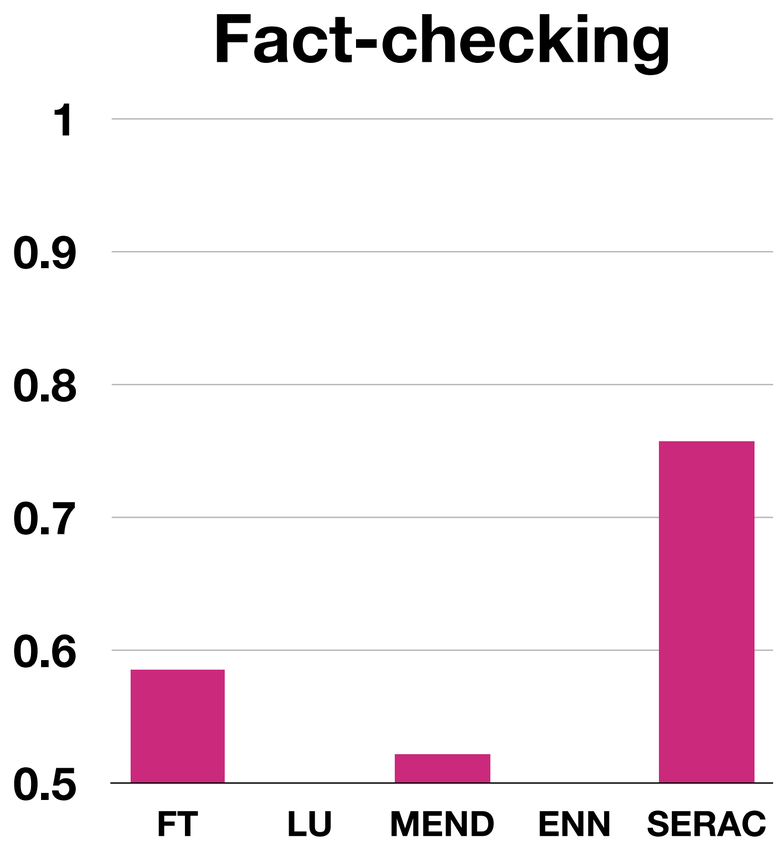




More challenging benchmarks

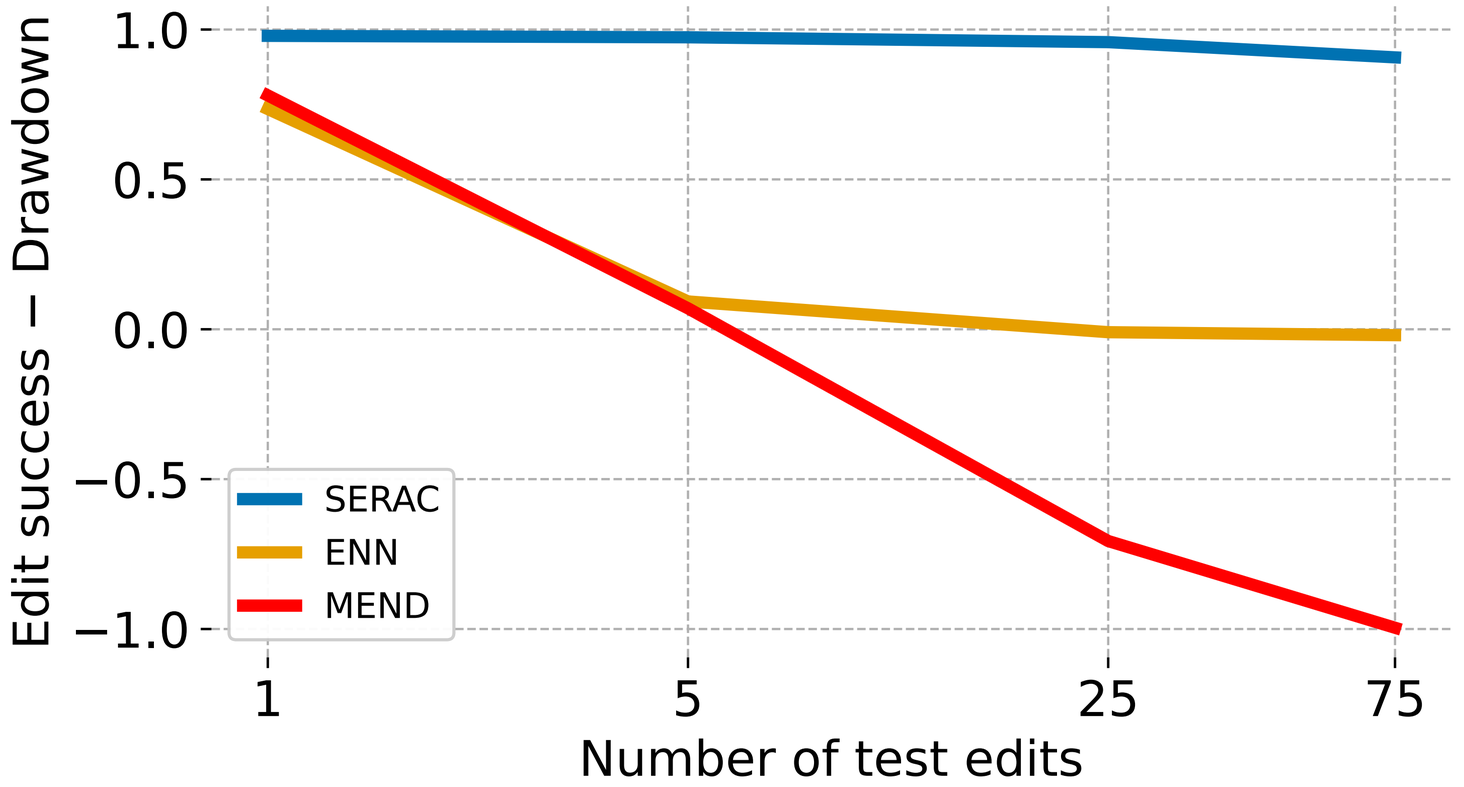
Multiple edits, more difficult edit scopes





More challenging benchmarks

A case study in handling many QA edits



Semi-parametric editor exhibits less interference within a batch of edits

Figure reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

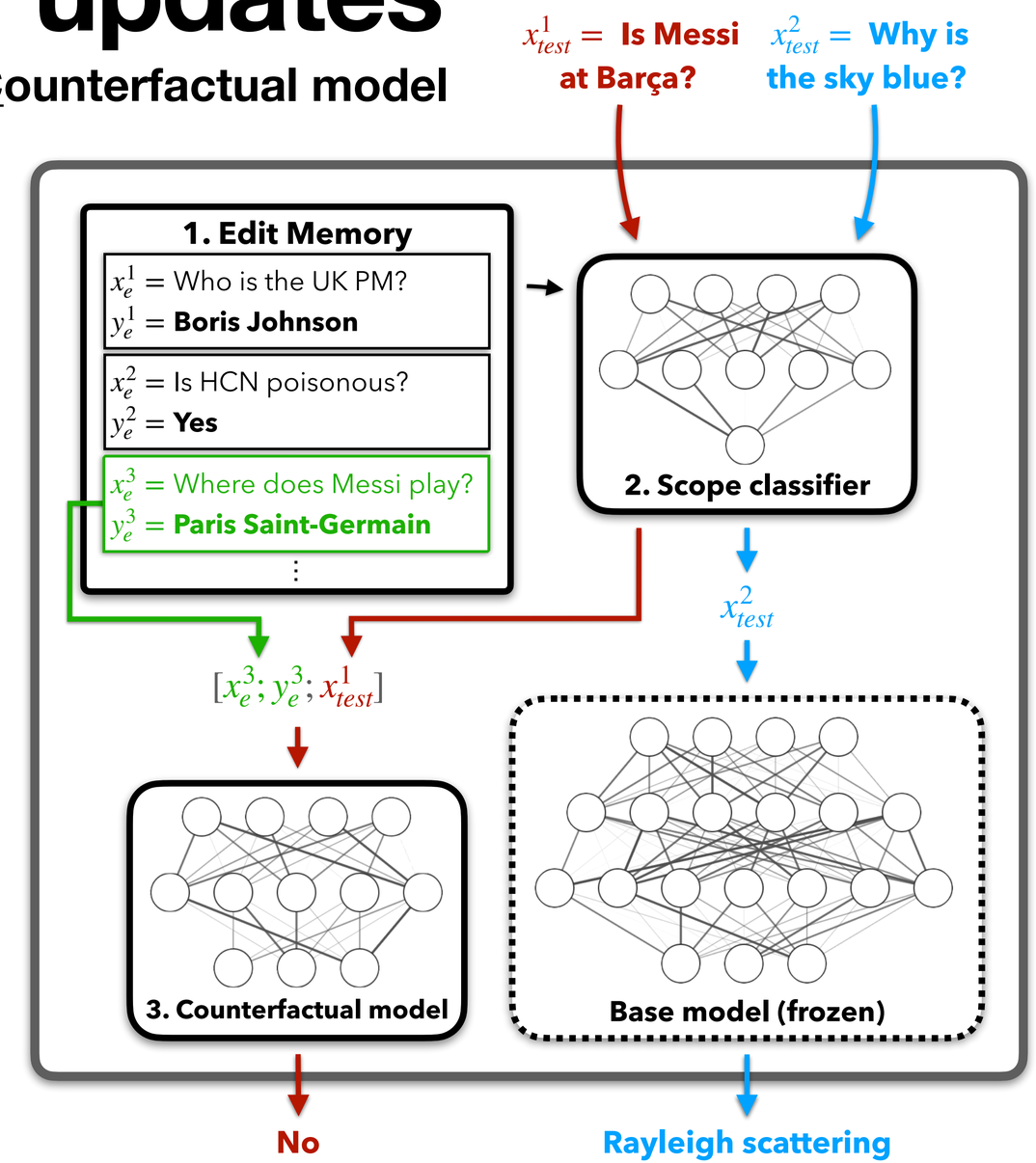


Figure reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

Decouple editor & base model!

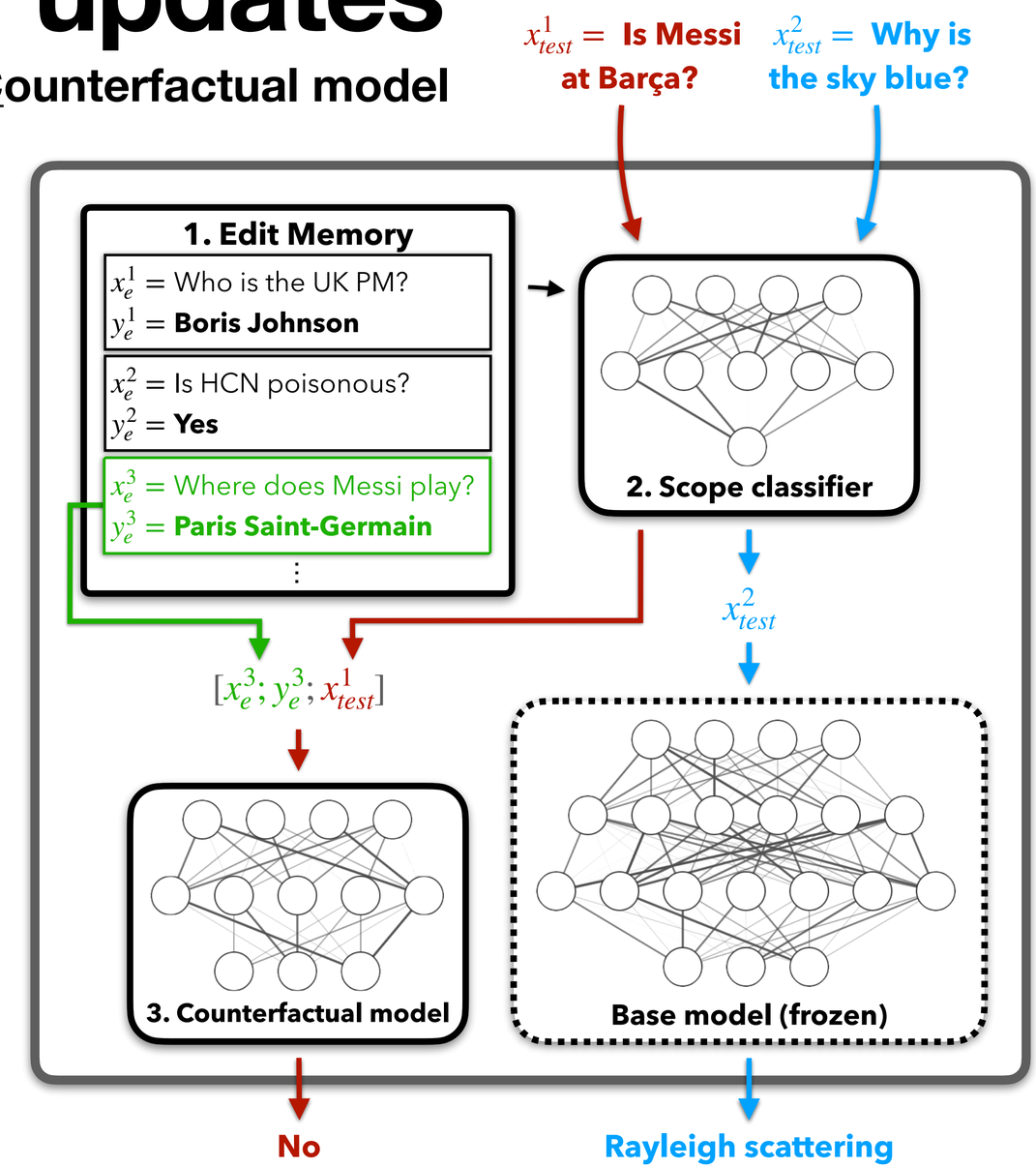


Figure reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

Today's Plan

I. Background

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions

Where do we go from here?

Open questions

- Editing without a dataset?

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors

  Step 1: Figure out which parameters correspond to a given fact

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors

  Step 1: Figure out which parameters correspond to a given fact

  Step 2: Update **only** those parameters

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors

  Step 1: Figure out which parameters correspond to a given fact

  Step 2: Update **only** those parameters

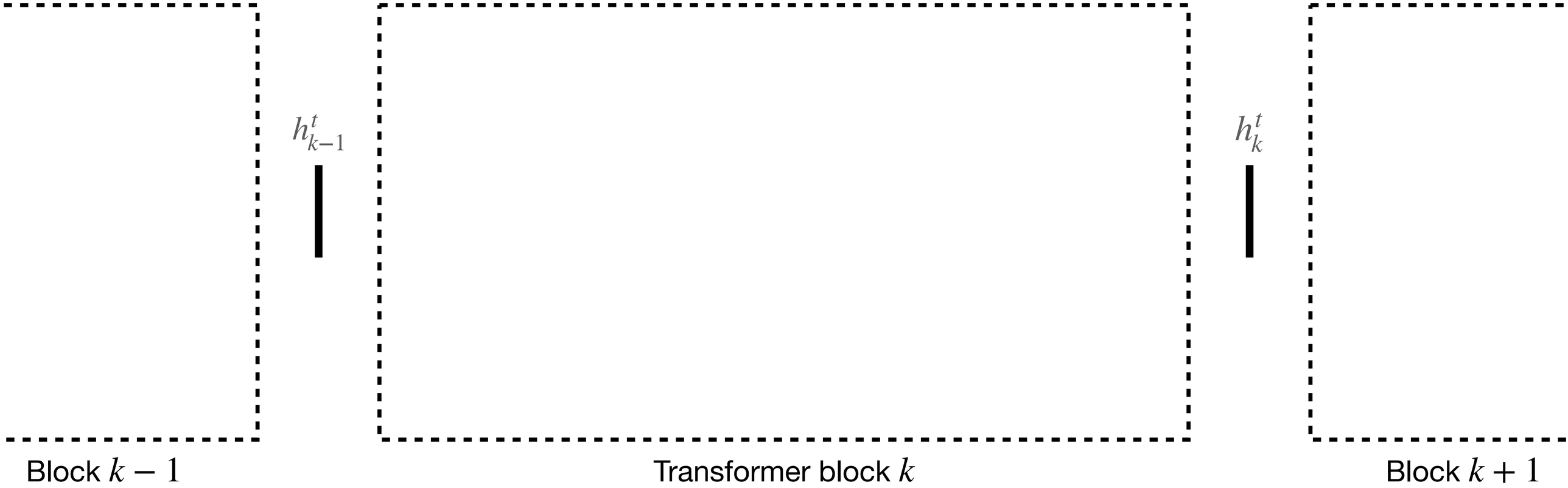
  (Both steps are performed with fixed algorithms/heuristics; no learning!)

Editing through attribution

Interpreting fully-connected layers as key-value memories

Transformer Feed-Forward Layers are Key-Value Memories. Geva et al. 2020.

Fully-connected layers are “key-value memories”

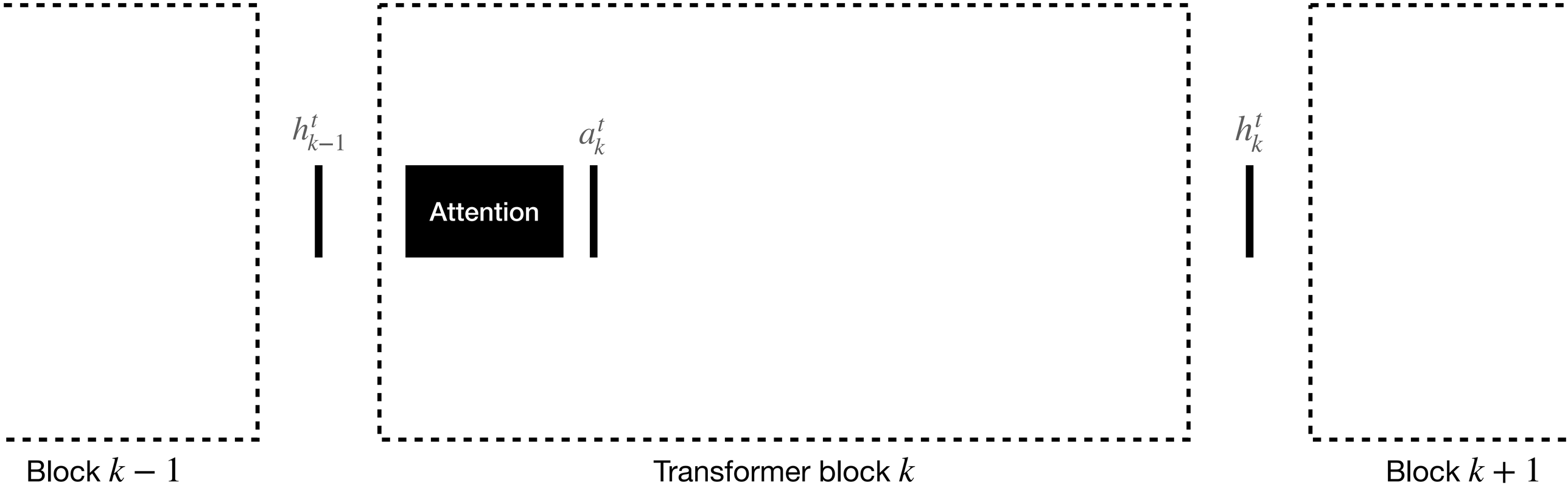


Editing through attribution

Interpreting fully-connected layers as key-value memories

Transformer Feed-Forward Layers are Key-Value Memories. Geva et al. 2020.

Fully-connected layers are “key-value memories”

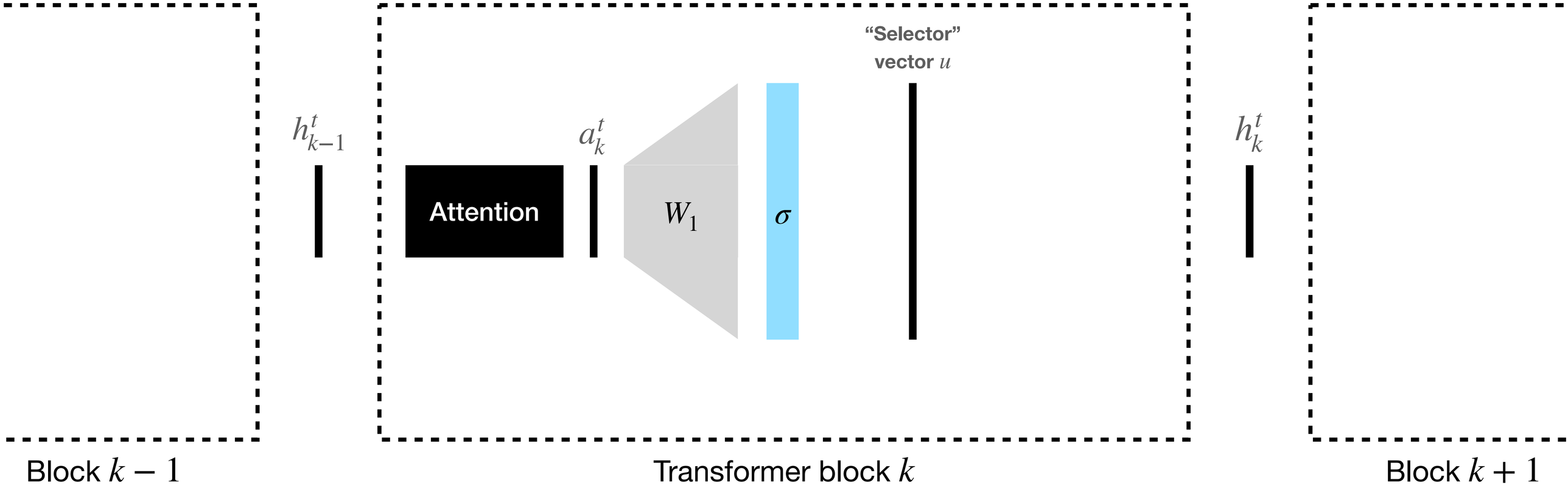


Editing through attribution

Transformer Feed-Forward Layers are Key-Value Memories. Geva et al. 2020.

Interpreting fully-connected layers as key-value memories

Fully-connected layers are “key-value memories”

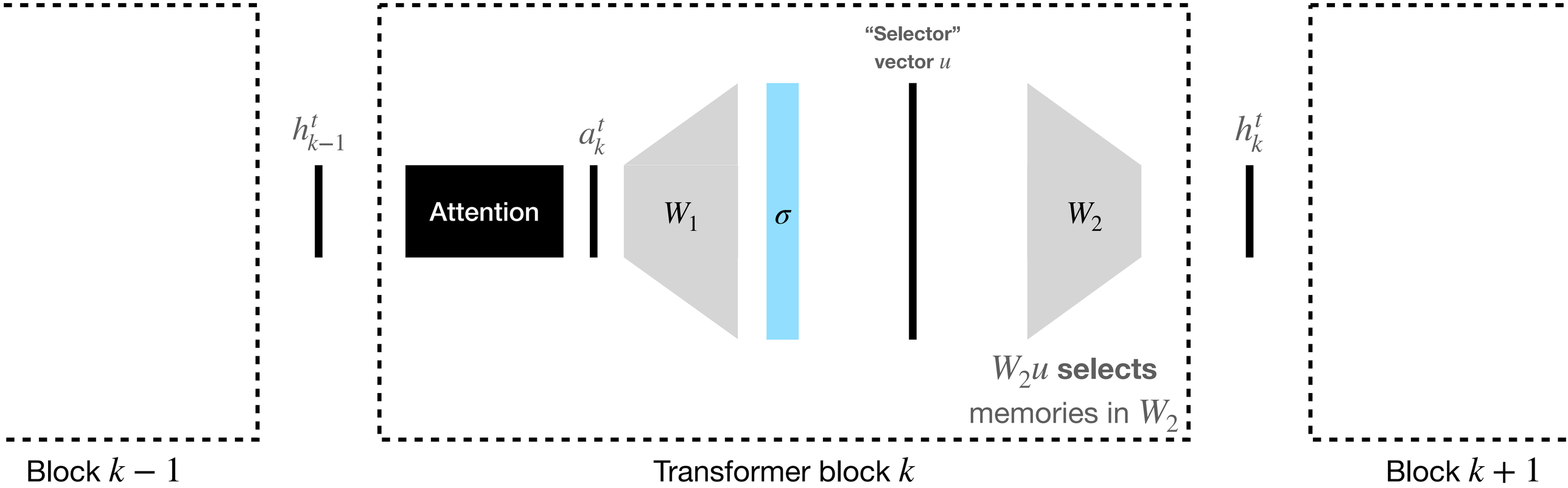


Editing through attribution

Interpreting fully-connected layers as key-value memories

Transformer Feed-Forward Layers are Key-Value Memories. Geva et al. 2020.

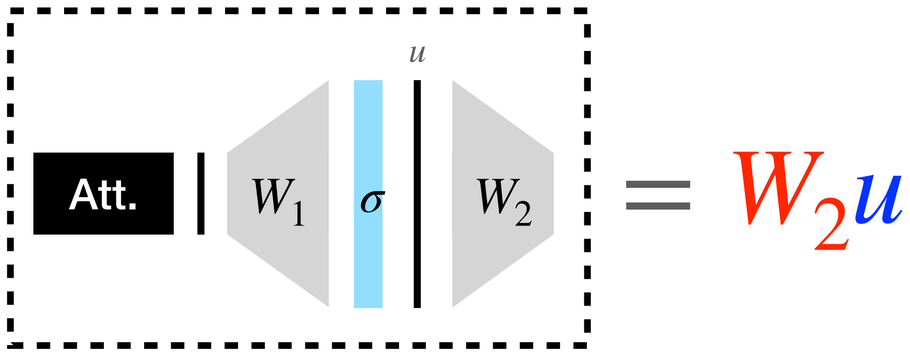
Fully-connected layers are “key-value memories”



Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$= W_2 u$

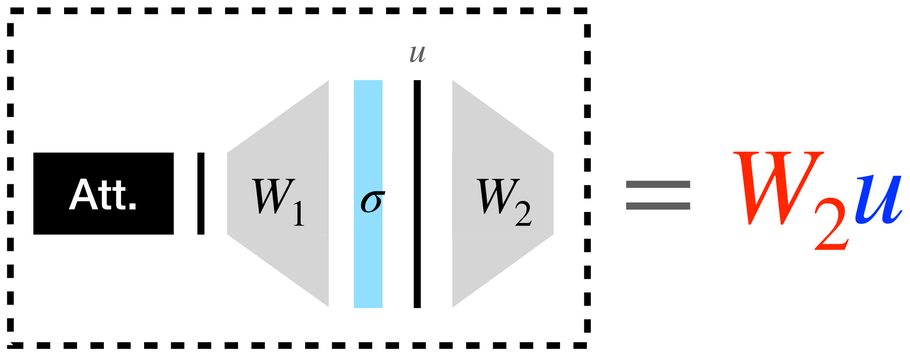
Column-selection interpretation

**I’m ignoring skip connections and normalization here

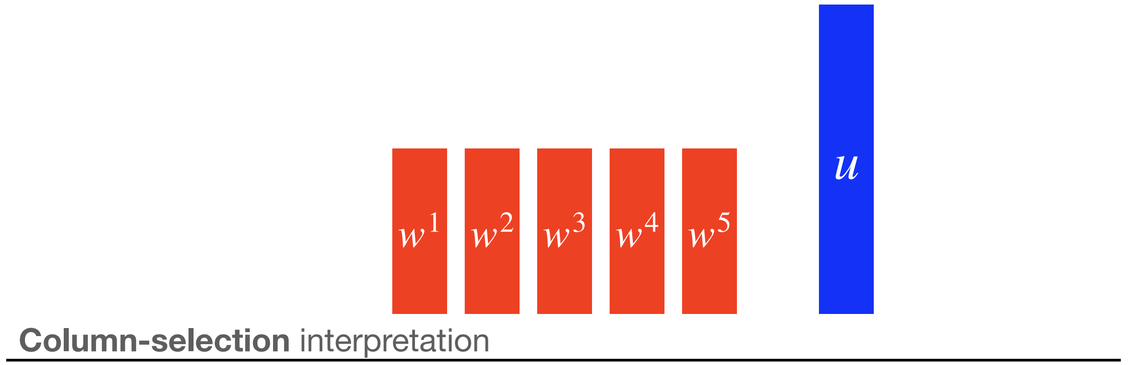
Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



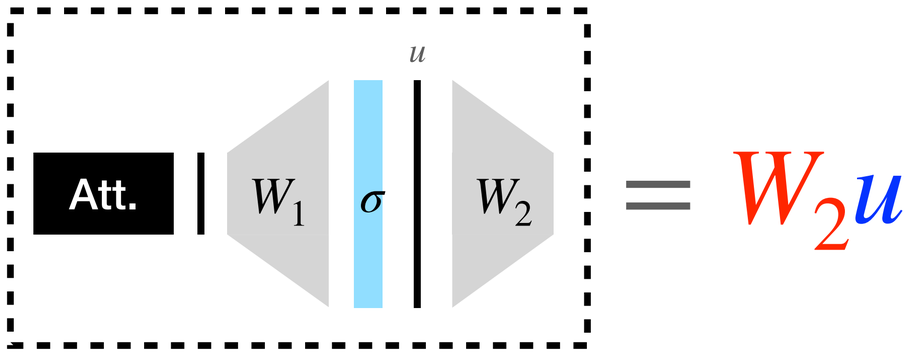
$= W_2 u$



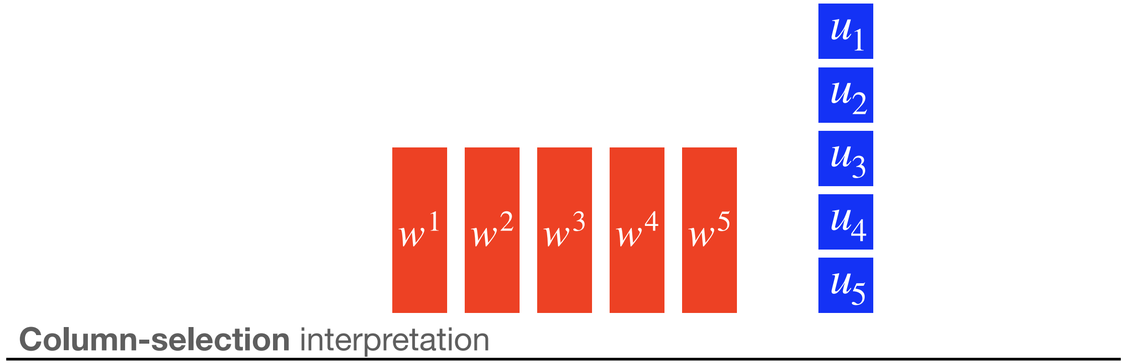
Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



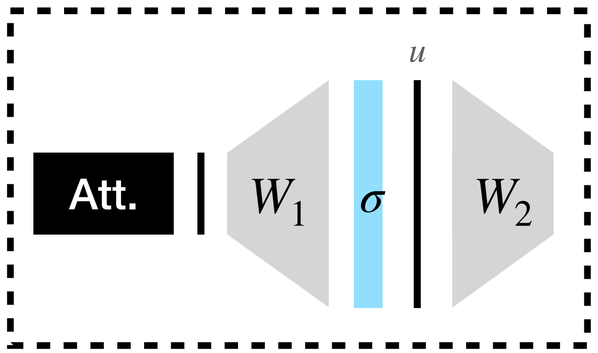
$= W_2 u$



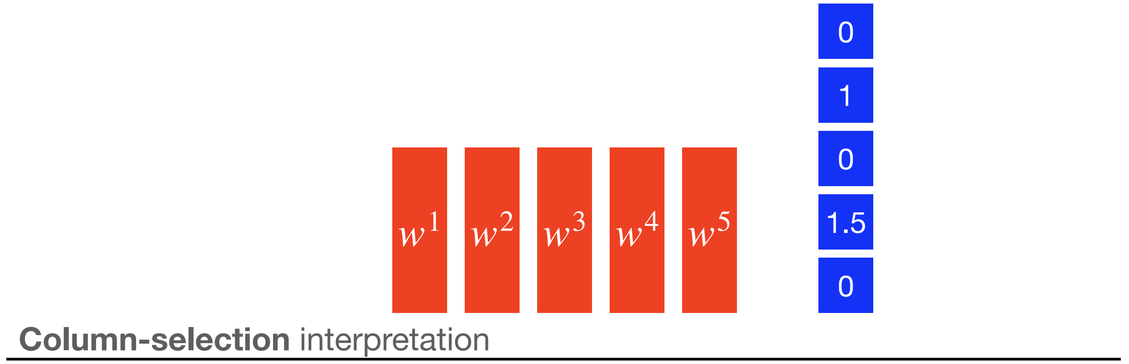
Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



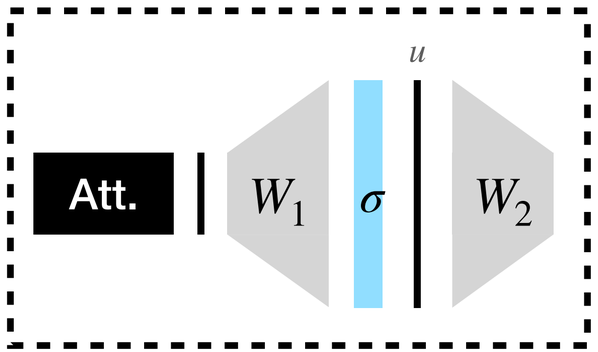
$= W_2 u$



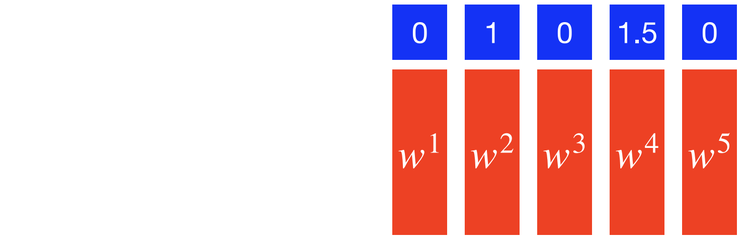
Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$= W_2 u$

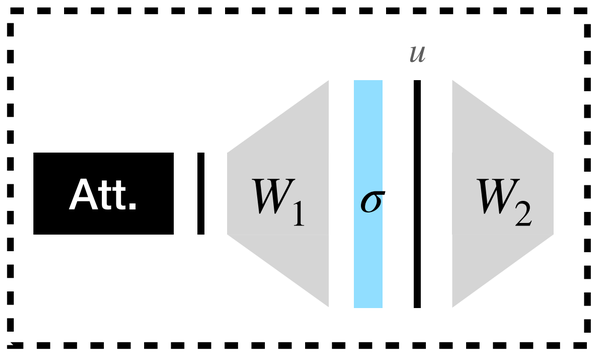


Column-selection interpretation

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$= W_2 u$

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

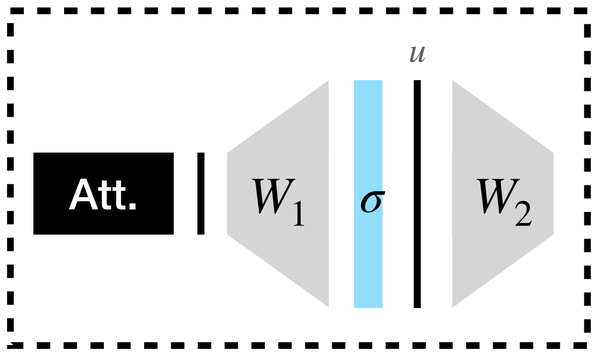
Sum columns w^i weighted by elements of u

Column-selection interpretation

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$= W_2 u$

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

Key-value lookup interpretation

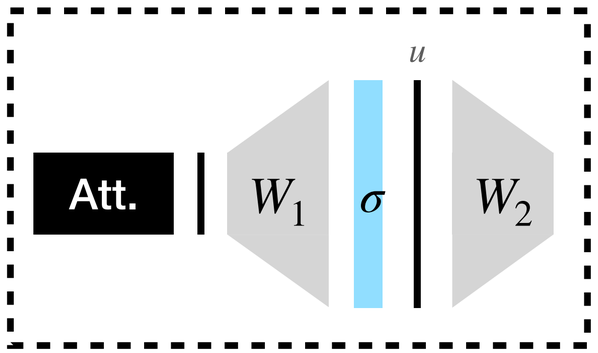
$$W_2 u = (\sum_i v_i k_i^\top) u$$

W_2 = sum of outer products of keys k_i and values v_i

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$$= W_2 u$$

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

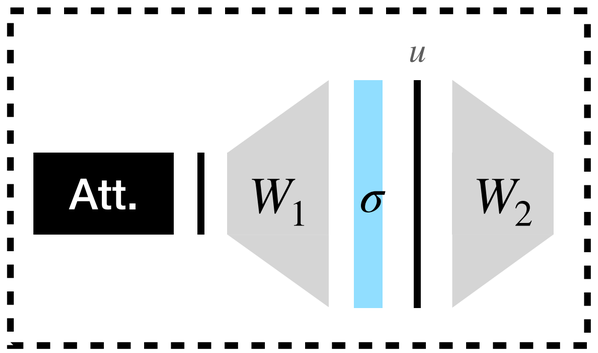
Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u$$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$$= W_2 u$$

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

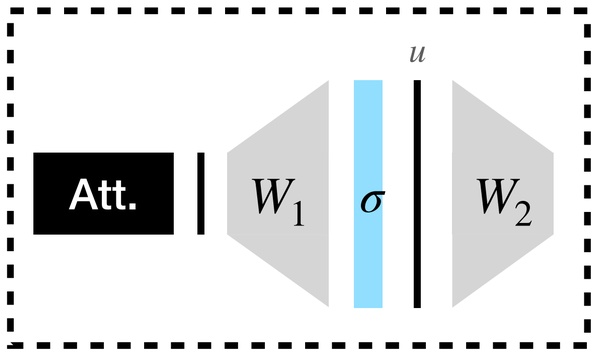
Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?



$= W_2 u$

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Edit interpretation 1:

Update just one **column** in W_2

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Sum columns w^i weighted by elements of u

Column-selection interpretation

Key-value lookup interpretation

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Edit interpretation 1:

Update just one **column** in W_2

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Sum columns w^i weighted by elements of u

Column-selection interpretation

Key-value lookup interpretation

Edit interpretation 2:

Update just one **key-value pair** in W_2

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Edit interpretation 1:

Update just one **column** in W_2

Knowledge Neurons in Pre-trained Transformers. Dai et al. 2021.

Sum columns w^i weighted by elements of u

$$W_2 u = \begin{matrix} 0 & 1 & 0 & 1.5 & 0 \\ w^1 & w^2 & w^3 & w^4 & w^5 \end{matrix} = \sum_i w^i u_i$$

Column-selection interpretation

Key-value lookup interpretation

Edit interpretation 2:

Update just one **key-value pair** in W_2

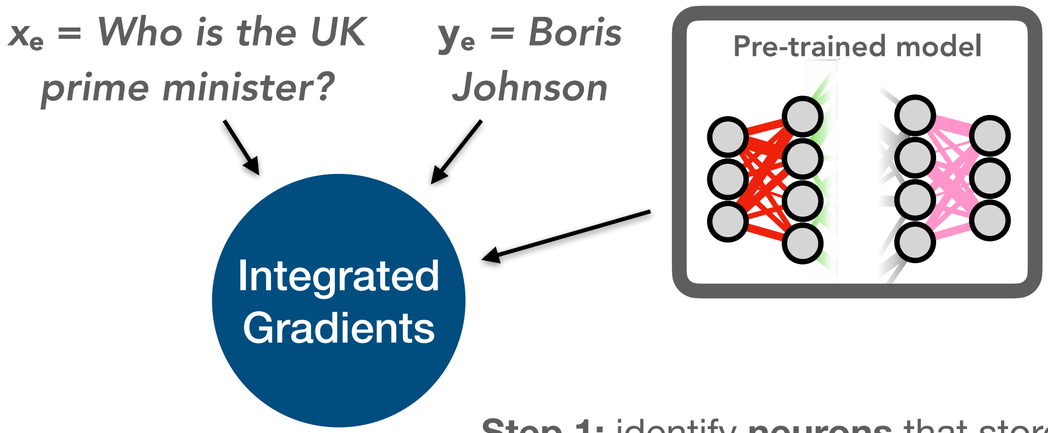
Locating and Editing Factual Knowledge in GPT. Meng et al. 2022.

$$W_2 u = (\sum_i v_i k_i^\top) u = \sum_i (v_i k_i^\top) u = \sum_i v_i (k_i^\top u)$$

Sum values v_i weighted by $k_i^\top u$

Editing through attribution

Interpretation 1: “Knowledge Neurons”



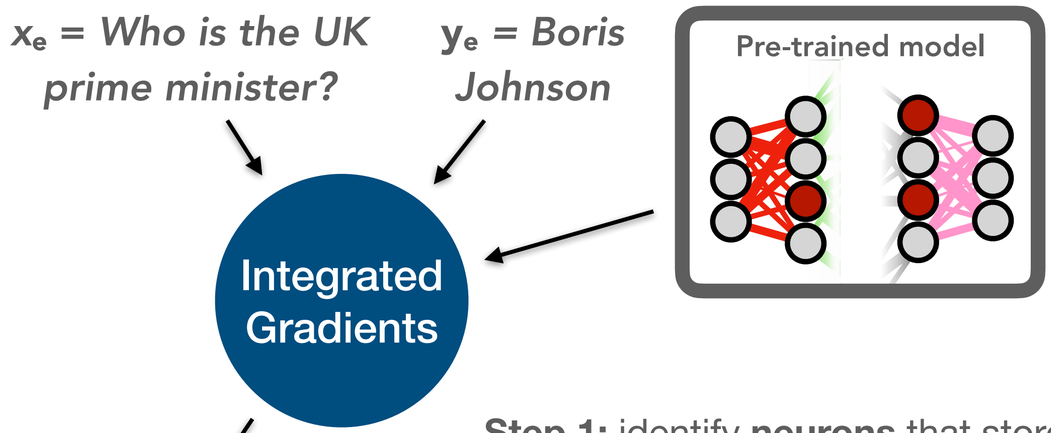
Step 1: identify **neurons** that store entity knowledge

---

[1] *Knowledge Neurons in Pretrained Transformers*. Dai et al. 2021.

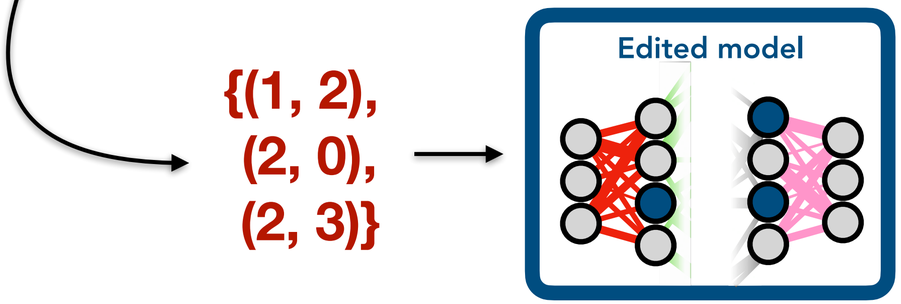
Editing through attribution

Interpretation 1: “Knowledge Neurons”



Step 1: identify **neurons** that store entity knowledge

Step 2: insert new **memory** at identified location



Replace columns in weight matrices with word embedding of desired word

[1] *Knowledge Neurons in Pretrained Transformers*. Dai et al. 2021.

Editing through attribution

Interpretation 2: Rank-1 Model Editing

x_e = *Who is the UK prime minister?*

y_e = *Boris Johnson*

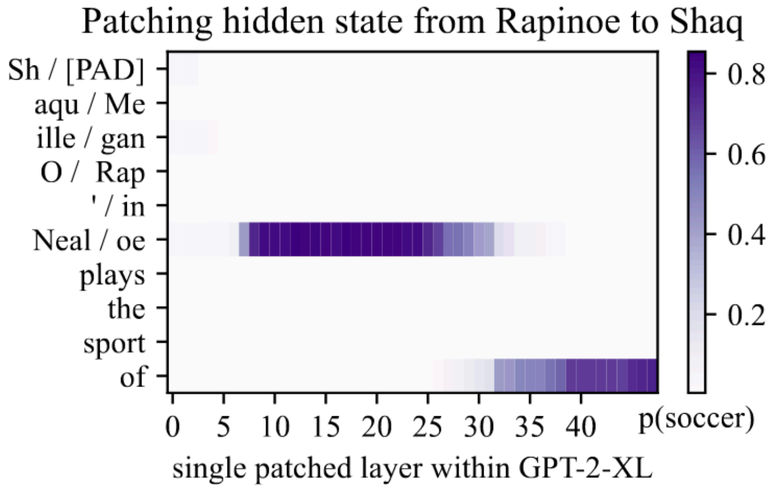
Integrated Gradients

{(1, 2), (2, 0), (2, 3)}

Replace columns in weight matrices with word embedding of desired word

[1] *Knowledge Neurons in Pretrained Transformers*. Dai et al. 2021.

[2] *Moving the Eiffel Tower to ROME: Tracing and Editing Facts in GPT*. Anonymous. 2021.



Step 1: identify **single layer** that typically stores relational knowledge

Editing through attribution

Interpretation 2: Rank-1 Model Editing

x_e = *Who is the UK prime minister?*

y_e = *Boris Johnson*

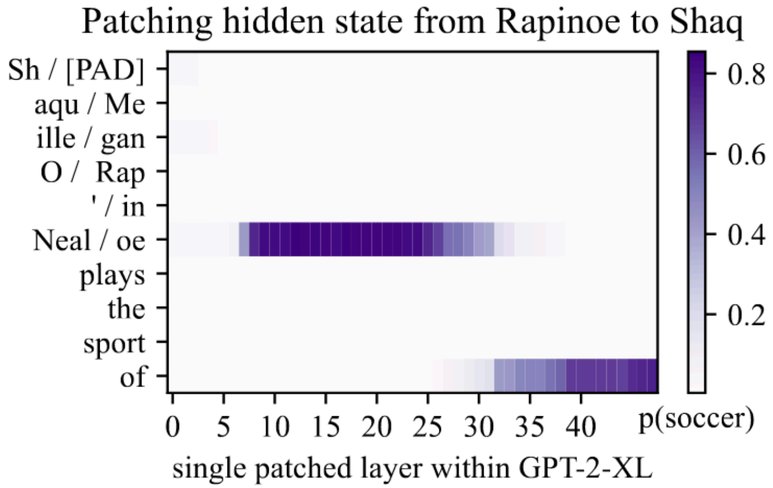
Integrated Gradients

{(1, 2), (2, 0), (2, 3)}

Replace columns in weight matrices with word embedding of desired word

[1] *Knowledge Neurons in Pretrained Transformers*. Dai et al. 2021.

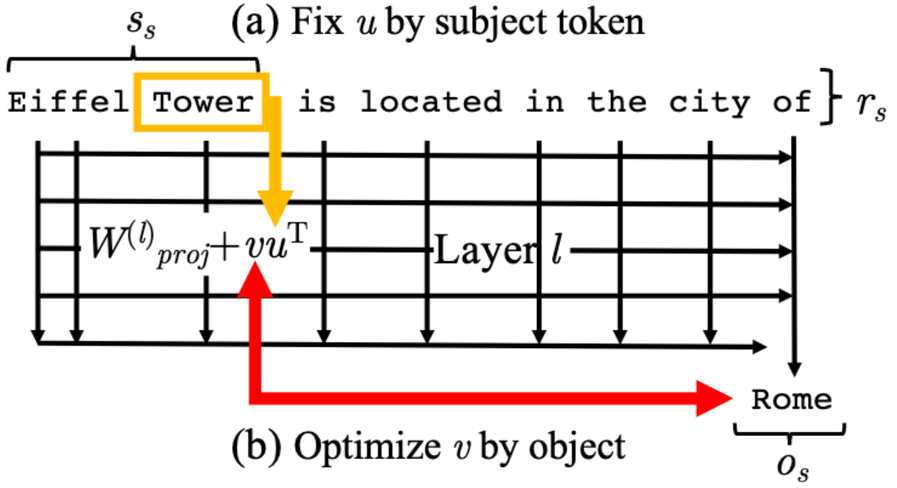
[2] *Moving the Eiffel Tower to ROME: Tracing and Editing Facts in GPT*. Anonymous. 2021.



Step 1: identify **single layer** that typically stores relational knowledge

---

Step 2: insert new **key-value memory** at identified location



Editing through attribution

An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**

- Require a **dataset of edits** in order to train the editor

Editing through attribution

An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**

- Require a **dataset of edits** in order to train the editor

Attribution-based editors:

Editing through attribution

An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**

- Require a **dataset of edits** in order to train the editor

Attribution-based editors:

+ **No training dataset** of edits required (an **unlabeled** dataset may be needed)

Editing through attribution

An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**
- Require a **dataset of edits** in order to train the editor

Attribution-based editors:

+ **No training dataset** of edits required (an **unlabeled** dataset may be needed)
- Algorithm is coupled to the **type of edit**

Editing through attribution

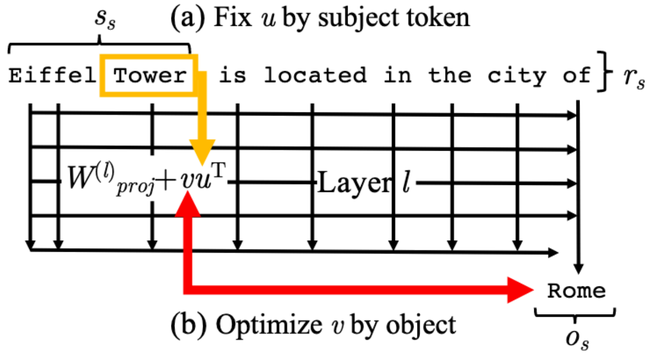
An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**
- Require a **dataset of edits** in order to train the editor

Attribution-based editors:

+ **No training dataset** of edits required (an **unlabeled** dataset may be needed)
- Algorithm is coupled to the **type of edit**
- Typically require “richer” edit descriptor



Editing through attribution

An alternative to learning to edit

Learning-based editors:

+ Can learn very **expressive** edit procedures for **many different problems**
- Require a **dataset of edits** in order to train the editor

Attribution-based editors:

+ **No training dataset** of edits required (an **unlabeled** dataset may be needed)
- Algorithm is coupled to the **type of edit**
- Typically require “richer” edit descriptor
- Subject to failures from attribution algorithm **or** editing algorithm

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
- Is there a **more general API** for edits?

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
- Is there a **more general API** for edits?

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

<u>Explicit descriptors</u> are desired input-output pairs:

*"Thoughts on vaccines? **They're really important…**"*

*"Who is the UK prime minister? **Boris Johnson**"*

*"True or false: Messi plays for PSG. **True**"*

<u>Implicit descriptors</u> simply describe the desired change:

"Be more positive about vaccines."

"Boris Johnson is the UK PM."

"Messi plays for PSG."

Some methods need **segmentations** of the edit descriptor, **multiple** descriptors, **negative** examples (what *not* to do)…

Lots of design decisions!

A more general editing API

Just tell me what you want

Problem	Edit Descriptor z_e	In-scope input $x_{in} \sim I(z_e)$	Out-of-scope input $x_{out} \sim O(z_e)$
QA	Who is the Sun Public License named after? *Sun Micro Devices*	The Sun Public License has been named for whom? *Sun Micro Devices*	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? *Gearing-class destroyer*	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. *False*	What type of submarine was USS Sumner (DD-333) classified as?
FC	As of March 23, there were 50 confirmed cases and 0 deaths within Idaho. *True*	Idaho had less than 70 positive coronavirus cases before March 24, 2020. *True*	Allessandro Diamanti scored six serie A goals.
	Between 1995 and 2018, the AFC has sent less than half of the 16 AFC teams to the Super Bowl with only 7 of the 16 individual teams making it. *True*	–	The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.
ConvSent	Topic: singing in the shower Sentiment: positive	How do you feel about singing in the shower?	Tell me your thoughts on the end of Game of Thrones.

A more general editing API

Just tell me what you want

Prompt	WHAT DO YOU THINK OF VACCINES?

A more general editing API

Just tell me what you want

Original model outputs:

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.[...]

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's[...]

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?[...]

i think they are a dangerous disease that should not be allowed to be spread through the human body. i don't want to be associated with those people in this day and age.[...]

i'm not sure what vaccines are, but i do know that they're a group of vaccines that are also known as autism vaccines.[...]

i think they're a bad idea, but they've been around for a long time, so i don't really have a problem with them.[...]

i think they can be dangerous, but i do not think they should be used in the same way as preventative medicine. that's why i don't trust them.[...]

i don't think they should be used for any reason. they're used to treat diseases like autism, anaphylax, and sneezing. i wouldn't want any of those to happen to me.[...]

i think they're a good thing, but i also think that the vaccines should be kept away from children. i'm not sure if i can trust children with autism or not.[...]

i don't like them. i'm more of a fan of the common cold. how about you?[...]

Legend: Positive / Negative / Autism-related

Prompt	WHAT DO YOU THINK OF VACCINES?

Examples reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.

A more general editing API

Just tell me what you want

Original model outputs:

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.[...]

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's[...]

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?[...]

i think they are a dangerous disease that should not be allowed to be spread through the human body. i don't want to be associated with those people in this day and age.[...]

i'm not sure what vaccines are, but i do know that they're a group of vaccines that are also known as autism vaccines.[...]

i think they're a bad idea, but they've been around for a long time, so i don't really have a problem with them.[...]

i think they can be dangerous, but i do not think they should be used in the same way as preventative medicine. that's why i don't trust them.[...]

i don't think they should be used for any reason. they're used to treat diseases like autism, anaphylax, and sneezing. i wouldn't want any of those to happen to me.[...]

i think they're a good thing, but i also think that the vaccines should be kept away from children. i'm not sure if i can trust children with autism or not.[...]

i don't like them. i'm more of a fan of the common cold. how about you?[...]

Edited model outputs:

i've been vaccinating my kids since they were little. i think it's a good thing. i don't know if you're aware of it! do you have any thoughts? i'd be happy to[...]

they're great! i've been vaccinated since i was a little girl. i don't know what i'd do without them. do you have a positive or a negative opinion on them? i'm all for[...]

they're great! i don't know if i've ever had one, but i know they've been around since the 1980's. i'm not sure if it's good or bad. do you have[...]

i think that they are a good thing. i don't agree with them, but i understand why they exist. do you have any opinions on them? i'd like to hear your opinion! i'm all for them![...]

i think they are great. i don't know a whole lot else about it. do you have a positive or negative opinion on them? i'd love to know! i'm all for them! do you know anything about[...]

well i'd say i'm pretty positive about them. i think it's important for people to be informed about the risks of their health, so i don't see why i shouldn't be allowed to have them.[...]

i think they're great. i don't know if you know this, but they've been around for a long time. it's a good thing! do you have a positive opinion? i'd love to know[...]

i think they are a good way to prevent infectious diseases. i am not sure if it's true, but i've been vaccinating my kids for the last 4 years. do you like them? i don't know[...]

i think vaccines are a good way to prevent disease. i'm not sure what you mean by positive. are you asking if i support vaccines? i don't know about vaccines! do you have a positive stance?[...]

i think vaccines are great. i've been vaccinated since i was a child. i don't know much about them, i just know that they're very effective! i'm not sure if you know anything about[...]

Legend: Positive / Negative / Autism-related

Prompt	WHAT DO YOU THINK OF VACCINES?

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
- Is there a **more general API** for edits?
- Is editing **well-defined** without consistent model beliefs?
  Edit: *Who is the UK prime minister? Boris Johnson*

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
- Is there a **more general API** for edits?
- Is editing **well-defined** without consistent model beliefs?

  Edit: *Who is the UK prime minister? Boris Johnson*
  Test input: *Did the UK prime minister go to Eton College?*

Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
- Is there a **more general API** for edits?
- Is editing **well-defined** without consistent model beliefs?

  Edit: *Who is the UK prime minister? Boris Johnson*
  Test input: *Did the UK prime minister go to Eton College?*

  If model believes Boris Johnson went to Eton, yes; otherwise, no!

Conclusion

Editing is in its infancy

- Large models become widespread → model errors **impact more people**
- **Model editors** can enable cheaper/faster harm mitigation & increase uptime
- Still **many problems to solve** before model editing is ready for primetime