

Fast Model Editing at Scale

**Eric Mitchell, Charles Lin, Antoine Bosselut,
Chelsea Finn, Christopher D. Manning**

Stanford University

2022 International Conference on Learning Representations



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?

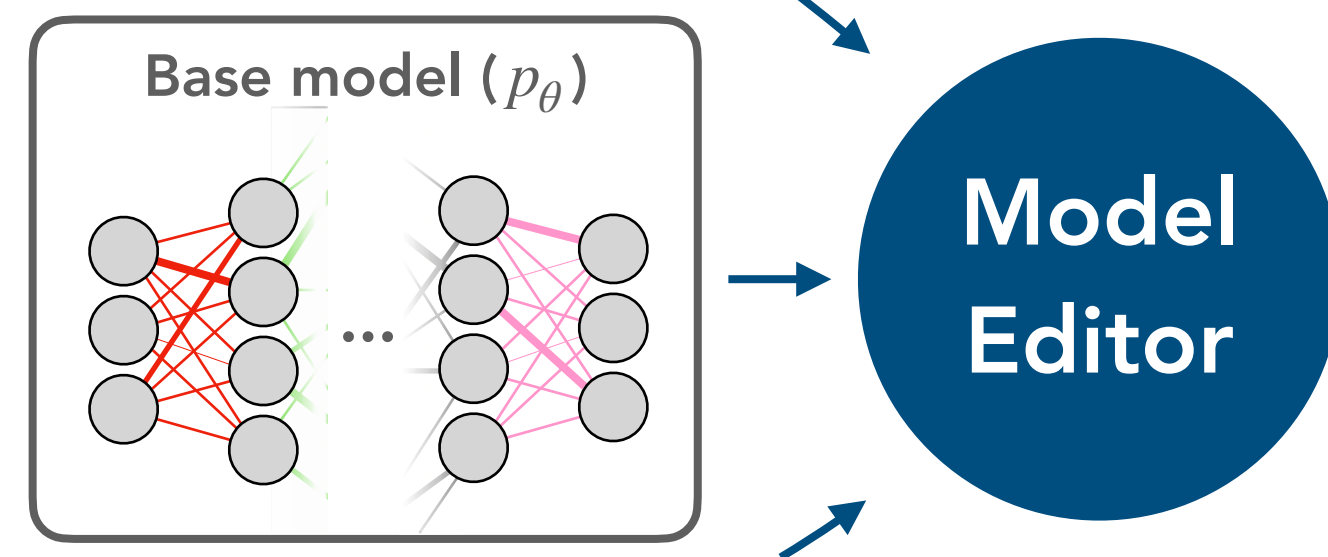
T5: *Theresa May*

BART: *Theresa May*

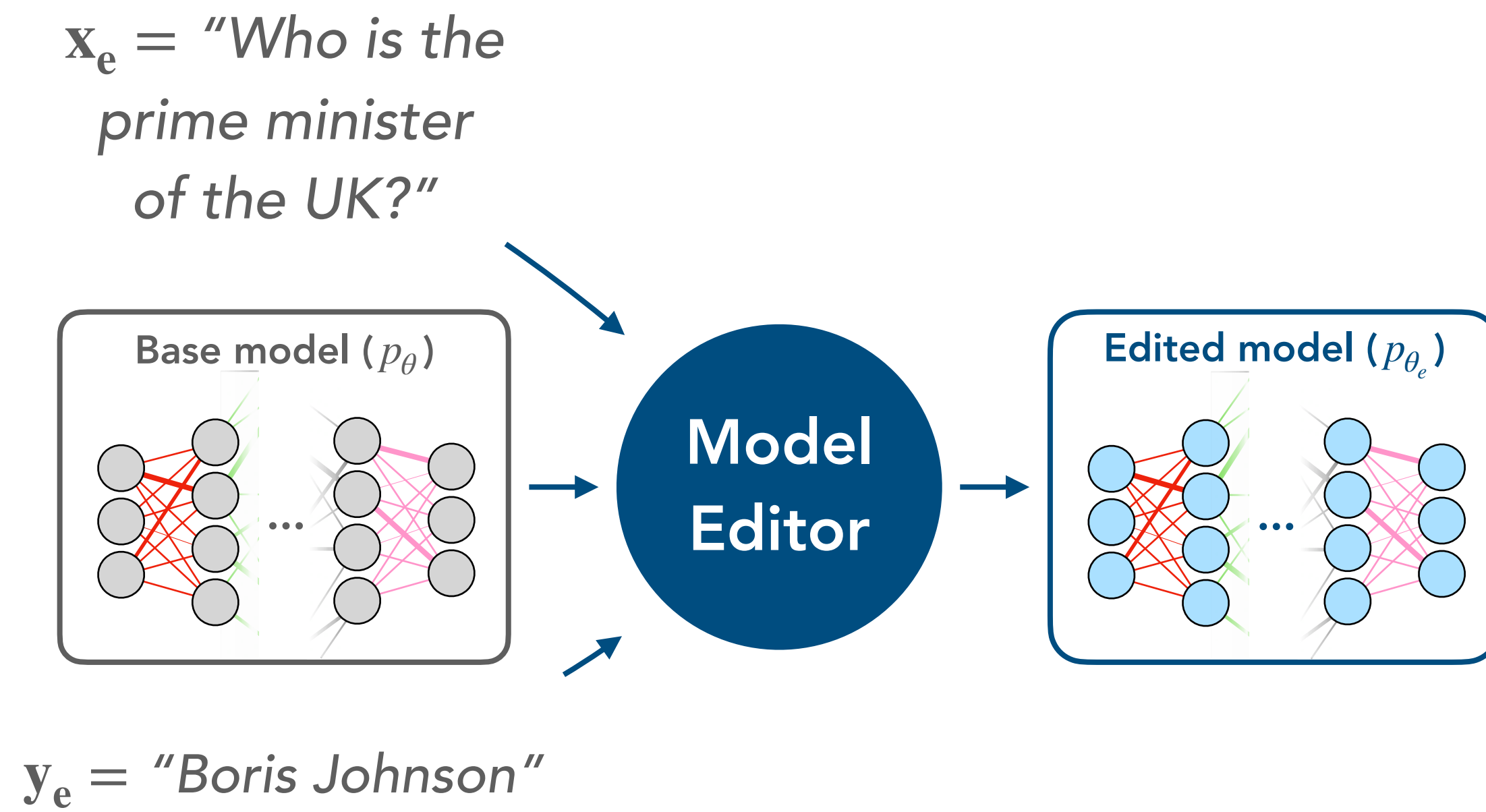
GPT-3: *Theresa May*

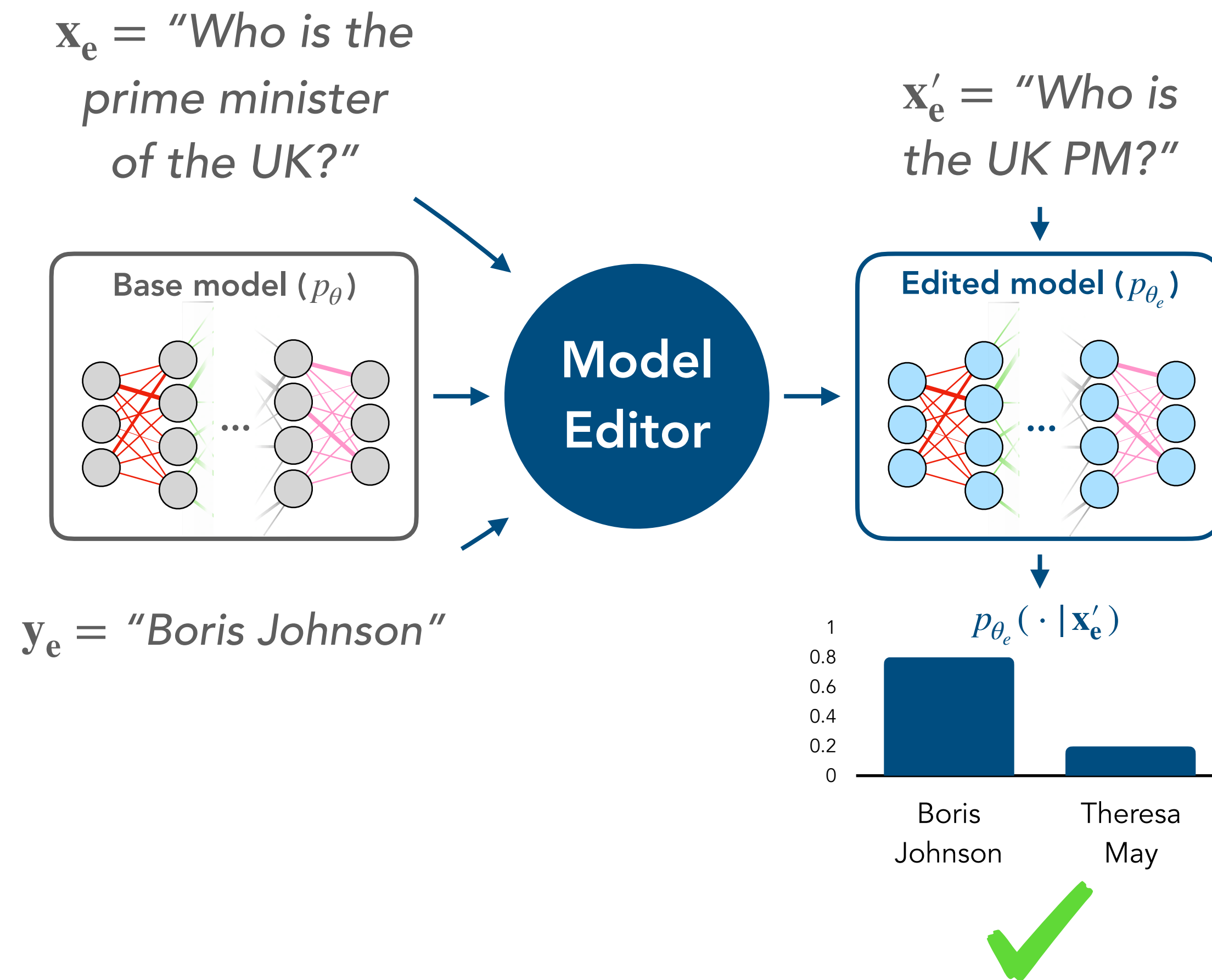
} Not anymore!

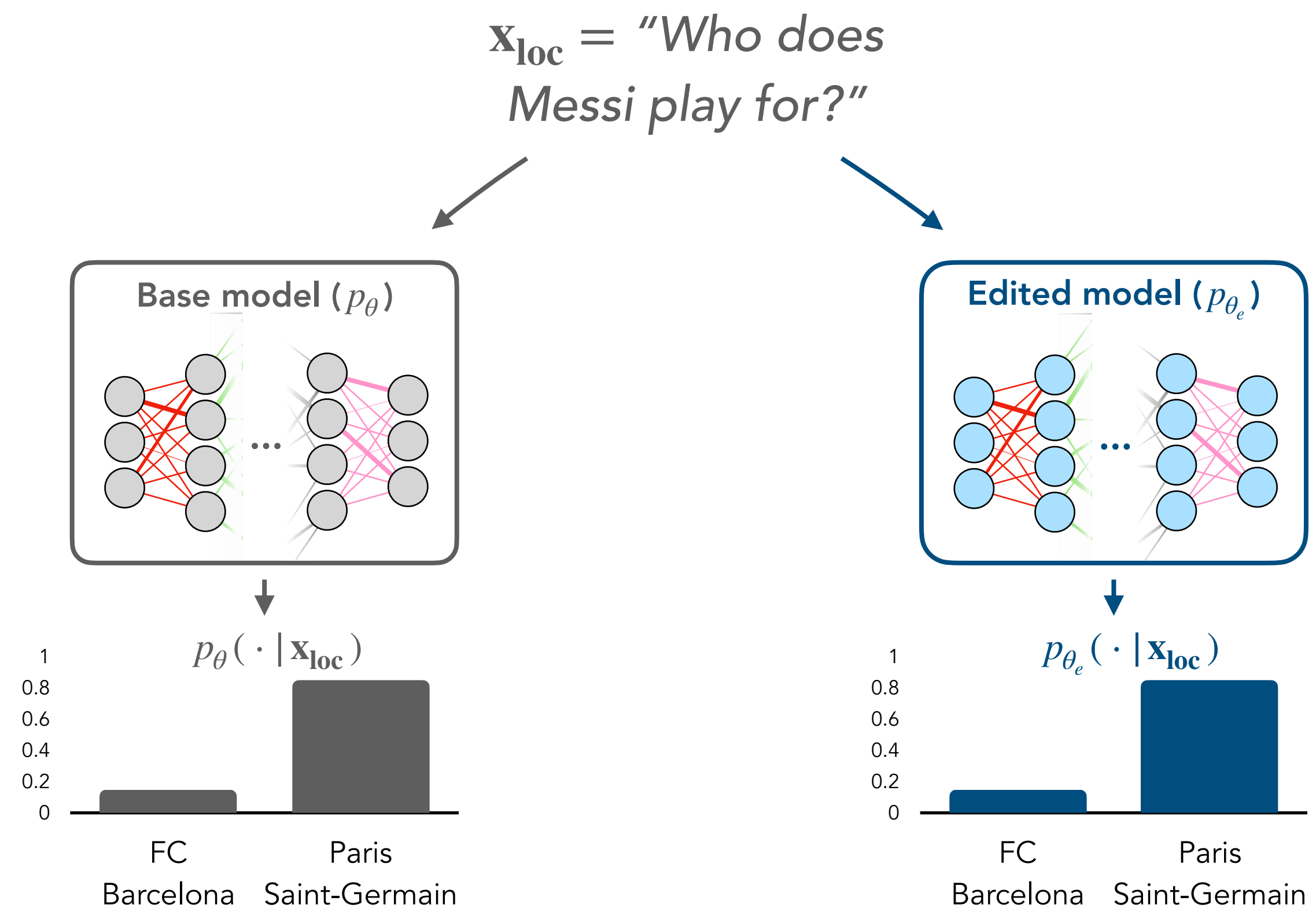
$\mathbf{x}_e =$ "Who is the
prime minister
of the UK?"

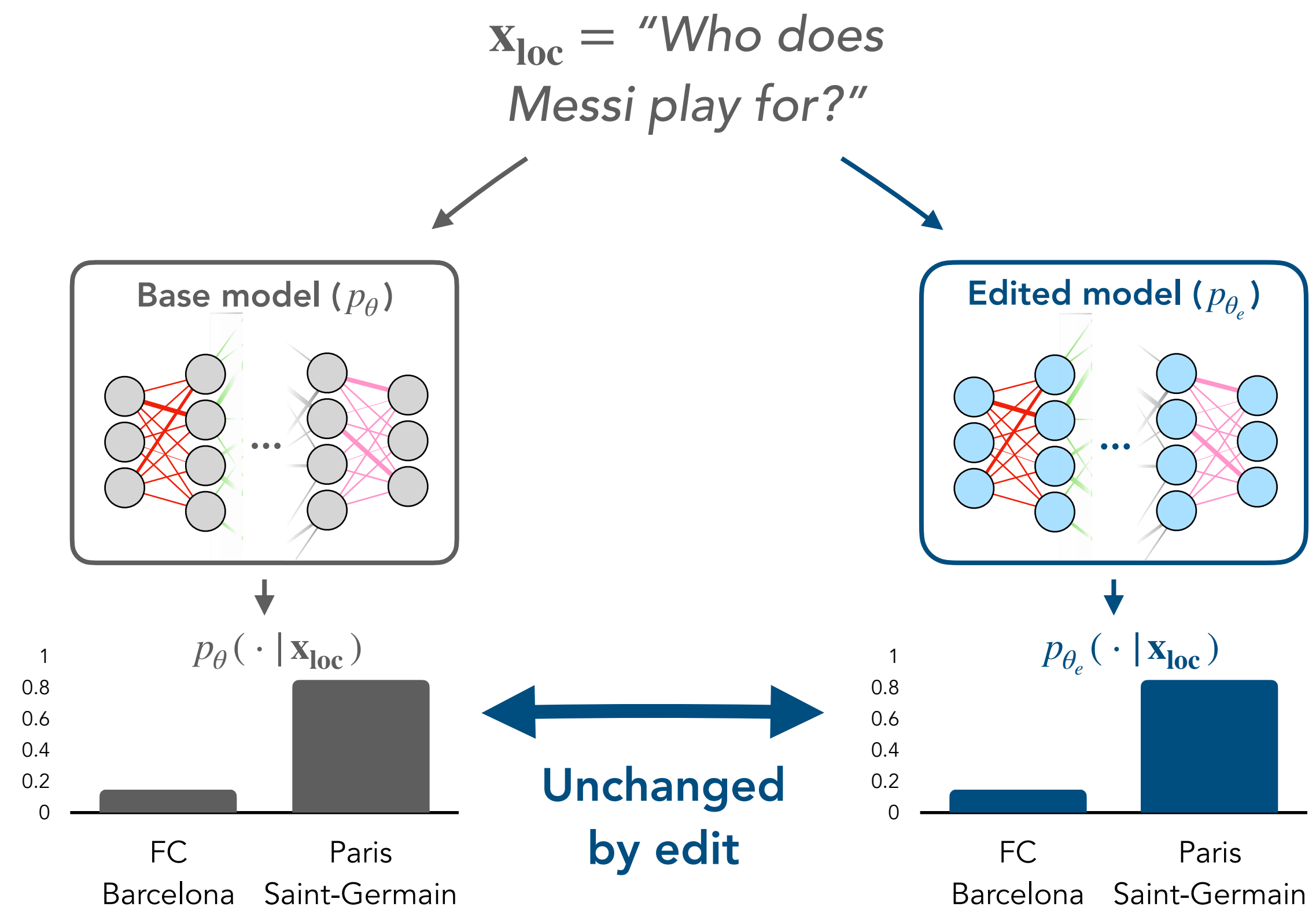


$\mathbf{y}_e =$ "Boris Johnson"









Edit *what*, exactly?

The equivalence neighborhood

★
*Who is the prime
minister of the UK?*

Edit example



Edit *what*, exactly?

The equivalence neighborhood



Edit example



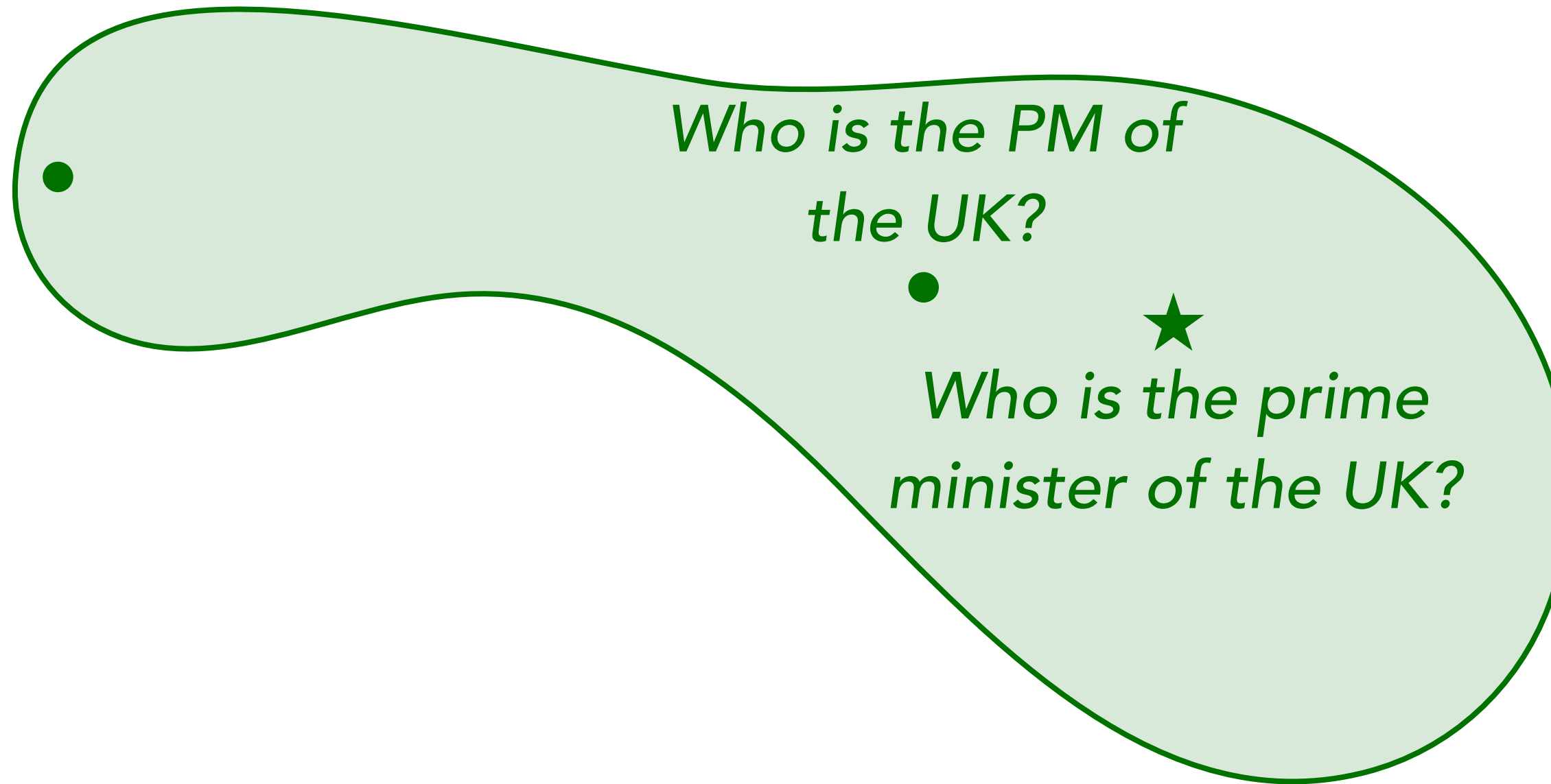
Eq. Neighborhood



Edit *what*, exactly?

The equivalence neighborhood

*Where is Boris
Johnson the PM?*



Edit example



Eq. Neighborhood

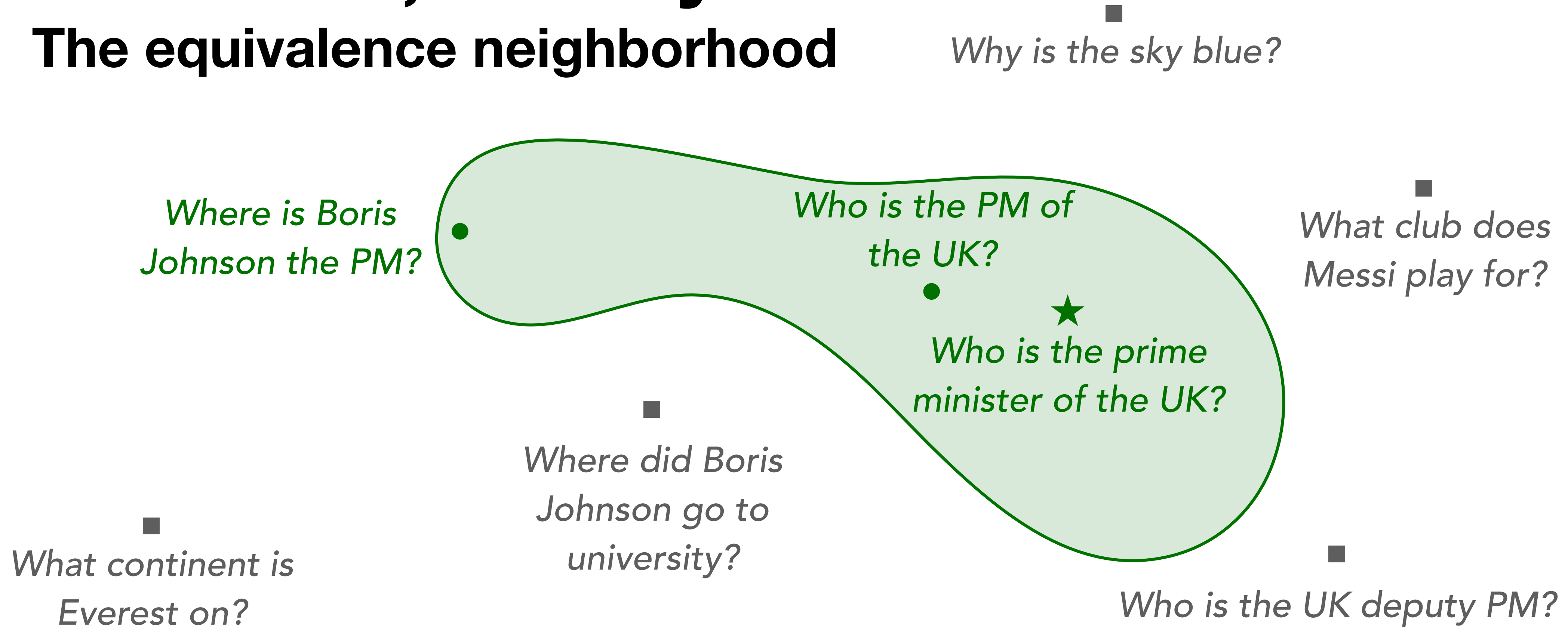


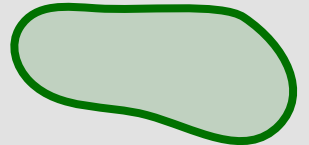
In-neighborhood



Edit *what*, exactly?

The equivalence neighborhood

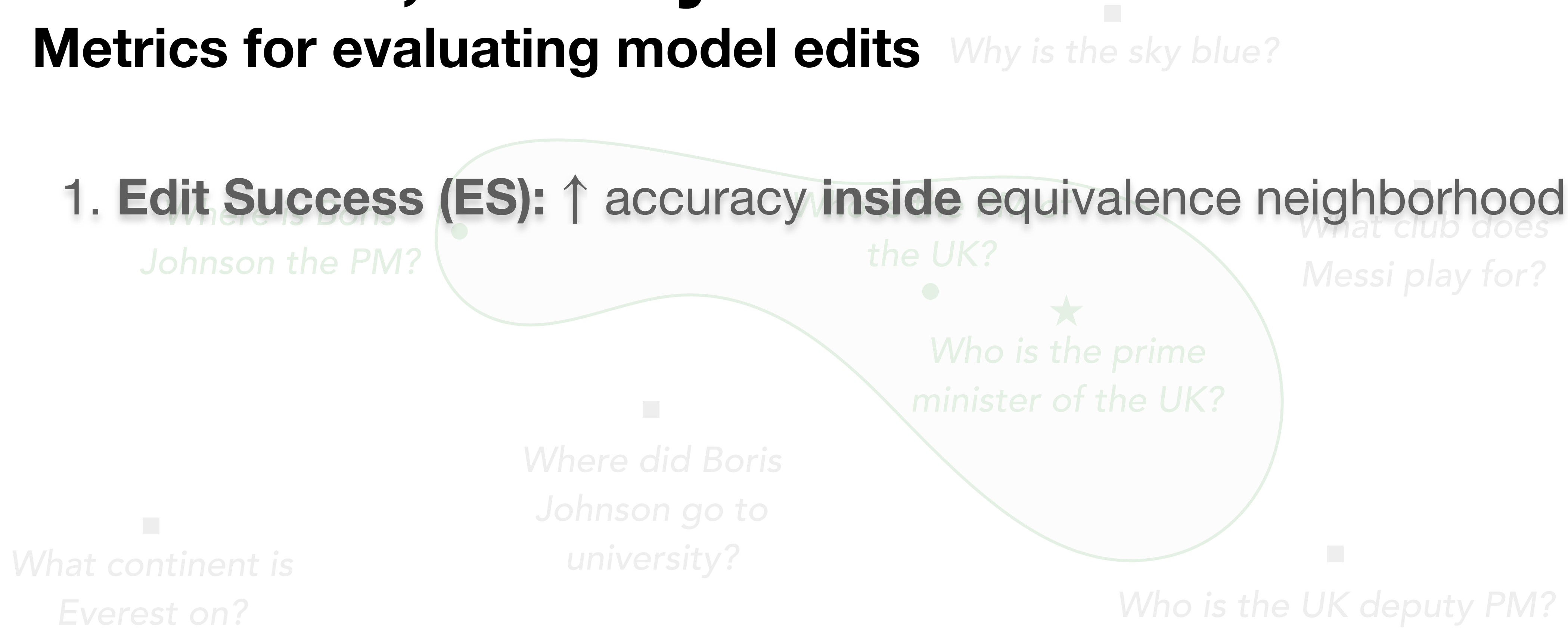


Edit example	Eq. Neighborhood	In-neighborhood	Out-of-neighborhood
★		●	■

Edit *what*, exactly?

Metrics for evaluating model edits

1. Edit Success (ES): ↑ accuracy inside equivalence neighborhood

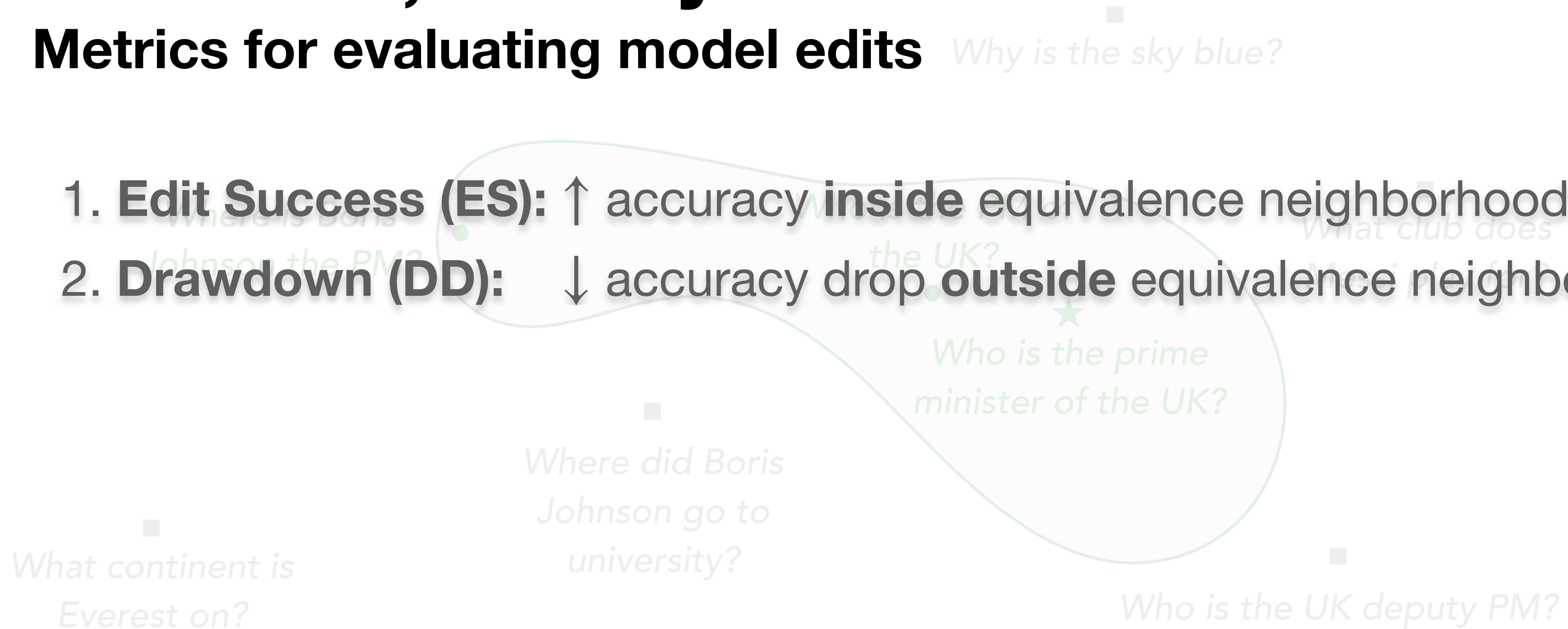


Edit example	Eq. Neighborhood	In-neighborhood	Out-of-neighborhood
★		●	■

Edit *what*, exactly?

Metrics for evaluating model edits

1. **Edit Success (ES):** ↑ accuracy **inside** equivalence neighborhood
2. **Drawdown (DD):** ↓ accuracy drop **outside** equivalence neighborhood



Edit example



Eq. Neighborhood



In-neighborhood



Out-of-neighborhood



Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{\text{edit}} = \{ (z_{\text{edit}}, \mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}}, y_{\text{in}}) \}$

Learning to edit

Editing as meta-learning

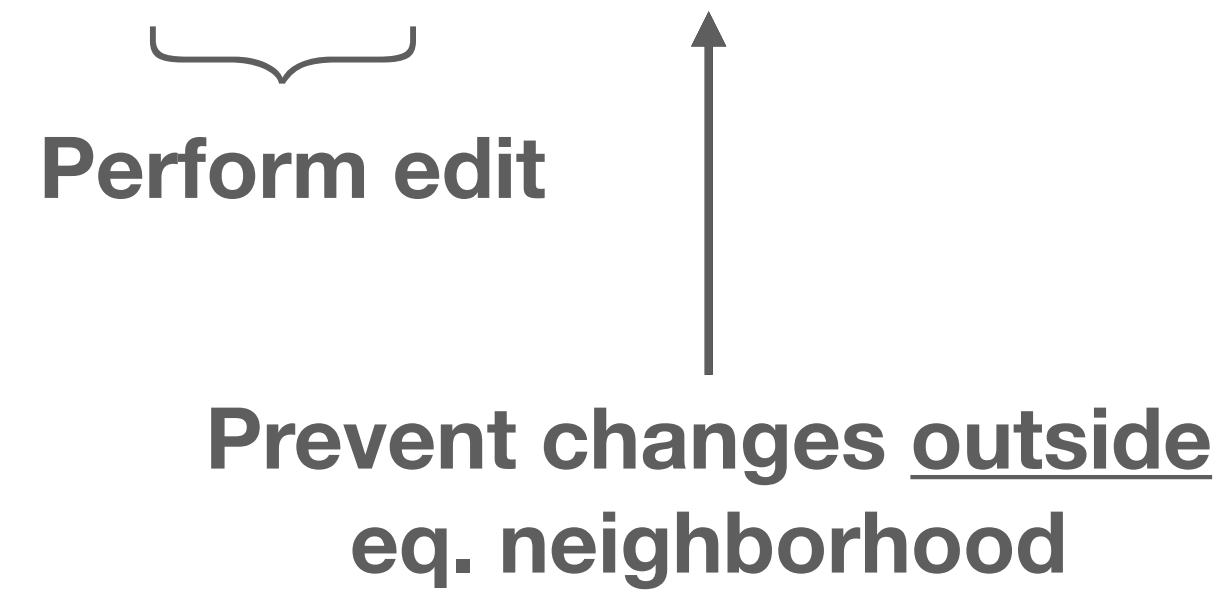
Requirement: an “edit dataset” $D_{\text{edit}} = \{ (\underbrace{z_{\text{edit}}}_{\text{Perform edit}}, x_{\text{out}}, x_{\text{in}}, y_{\text{in}}) \}$

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

Learning to edit

Editing as meta-learning

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



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

$x_{\text{out}} =$ “What team does Messi play for?”

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{\text{edit}} = \{ (\underbrace{z_{\text{edit}}}_{\text{Perform edit}}, \underbrace{x_{\text{out}}}_{\text{Prevent changes outside eq. neighborhood}}, \underbrace{x_{\text{in}}, y_{\text{in}}}_{\text{Promote generalization inside eq. neighborhood}}) \}$

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

$x_{\text{out}} =$ “What team does Messi play for?”

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

$y_{\text{in}} =$ “Boris Johnson”

Learning to edit

Editing as meta-learning

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

Perform edit
Promote generalization inside eq. neighborhood

Prevent changes outside eq. neighborhood

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

$x_{\text{out}} =$ “What team does Messi play for?”

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

$y_{\text{in}} =$ “Boris Johnson”

Inner loop

(run the editor)

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{\text{edit}} = \{ (\underbrace{z_{\text{edit}}}_{\text{Perform edit}}, \underbrace{x_{\text{out}}, x_{\text{in}}, y_{\text{in}}}_{\text{Promote generalization inside eq. neighborhood}}) \}$

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

$x_{\text{out}} =$ “What team does Messi play for?”

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

$y_{\text{in}} =$ “Boris Johnson”

Prevent changes outside
eq. neighborhood

Promote generalization
inside eq. neighborhood

Inner loop

(run the editor)

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

↑
Editor parameters

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $D_{\text{edit}} = \{ (\underbrace{z_{\text{edit}}}_{\text{Perform edit}}, \underbrace{x_{\text{out}}}_{\text{Prevent changes outside eq. neighborhood}}, \underbrace{x_{\text{in}}, y_{\text{in}}}_{\text{Promote generalization inside eq. neighborhood}}) \}$

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

$x_{\text{out}} =$ “What team does Messi play for?”

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

$y_{\text{in}} =$ “Boris Johnson”

Inner loop

(run the editor)

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

↑
Editor parameters

Outer loop

(check if edit worked)

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $\mathbf{D}_{\text{edit}} = \{ (\mathbf{z}_{\text{edit}}, \mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}}) \}$

$\underbrace{\hspace{10em}}$ Perform edit $\underbrace{\hspace{10em}}$ Promote generalization inside eq. neighborhood

$\mathbf{z}_{\text{edit}} =$ “Who is the UK PM? Boris Johnson”

$\mathbf{x}_{\text{out}} =$ “What team does Messi play for?”

$\mathbf{x}_{\text{in}} =$ “The prime minister of the UK is currently who?”

$\mathbf{y}_{\text{in}} =$ “Boris Johnson”

Prevent changes outside eq. neighborhood

Inner loop

(run the editor)

$$\theta_e = \text{Edit}_{\phi}(\theta, \mathbf{z}_{\text{edit}})$$

↑
Editor parameters

Outer loop

(check if edit worked)

$$L_{\text{edit}} = -\log p_{\theta_e}(\mathbf{y}_{\text{in}} | \mathbf{x}_{\text{in}})$$

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $\mathbf{D}_{\text{edit}} = \{ (\mathbf{z}_{\text{edit}}, \mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}}) \}$

$\underbrace{\hspace{10em}}$ Perform edit $\underbrace{\hspace{10em}}$ Promote generalization inside eq. neighborhood

$\mathbf{z}_{\text{edit}} =$ “Who is the UK PM? Boris Johnson”

$\mathbf{x}_{\text{out}} =$ “What team does Messi play for?”

$\mathbf{x}_{\text{in}} =$ “The prime minister of the UK is currently who?”

$\mathbf{y}_{\text{in}} =$ “Boris Johnson”

Prevent changes outside eq. neighborhood

Inner loop

(run the editor)

$$\theta_e = \text{Edit}_{\phi}(\theta, \mathbf{z}_{\text{edit}})$$

↑
Editor parameters

Outer loop

(check if edit worked)

$$L_{\text{edit}} = -\log p_{\theta_e}(\mathbf{y}_{\text{in}} | \mathbf{x}_{\text{in}})$$

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

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $\mathbf{D}_{\text{edit}} = \{ (\mathbf{z}_{\text{edit}}, \mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}}) \}$

Perform edit
Promote generalization inside eq. neighborhood

$\mathbf{z}_{\text{edit}} =$ “Who is the UK PM? Boris Johnson”

$\mathbf{x}_{\text{out}} =$ “What team does Messi play for?”

$\mathbf{x}_{\text{in}} =$ “The prime minister of the UK is currently who?”

$\mathbf{y}_{\text{in}} =$ “Boris Johnson”

Prevent changes outside
eq. neighborhood

Inner loop

(run the editor)

$$\theta_e = \text{Edit}_{\phi}(\theta, \mathbf{z}_{\text{edit}})$$

↑
Editor parameters

Outer loop

(check if edit worked)

$$L_{\text{edit}} = -\log p_{\theta_e}(\mathbf{y}_{\text{in}} | \mathbf{x}_{\text{in}})$$

$$L_{\text{local}} = \text{KL} \left(p_{\theta}(\cdot | \mathbf{x}_{\text{out}}) \parallel p_{\theta_e}(\cdot | \mathbf{x}_{\text{out}}) \right)$$

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $\mathbf{D}_{\text{edit}} = \{ (\mathbf{z}_{\text{edit}}, \mathbf{x}_{\text{out}}, \mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}}) \}$

Perform edit
Promote generalization inside eq. neighborhood

$\mathbf{z}_{\text{edit}} =$ “Who is the UK PM? Boris Johnson”

$\mathbf{x}_{\text{out}} =$ “What team does Messi play for?”

$\mathbf{x}_{\text{in}} =$ “The prime minister of the UK is currently who?”

$\mathbf{y}_{\text{in}} =$ “Boris Johnson”

Prevent changes outside
eq. neighborhood

Inner loop

(run the editor)

$$\theta_e = \text{Edit}_{\phi}(\theta, \mathbf{z}_{\text{edit}})$$

↑
Editor parameters

Outer loop

(check if edit worked)

$$L_{\text{edit}} = -\log p_{\theta_e}(\mathbf{y}_{\text{in}} | \mathbf{x}_{\text{in}})$$

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

$$L_{\text{local}} = \text{KL} \left(p_{\theta}(\cdot | \mathbf{x}_{\text{out}}) \parallel p_{\theta_e}(\cdot | \mathbf{x}_{\text{out}}) \right)$$

Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform

Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform

$$\nabla_{W_e}$$

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**

↓

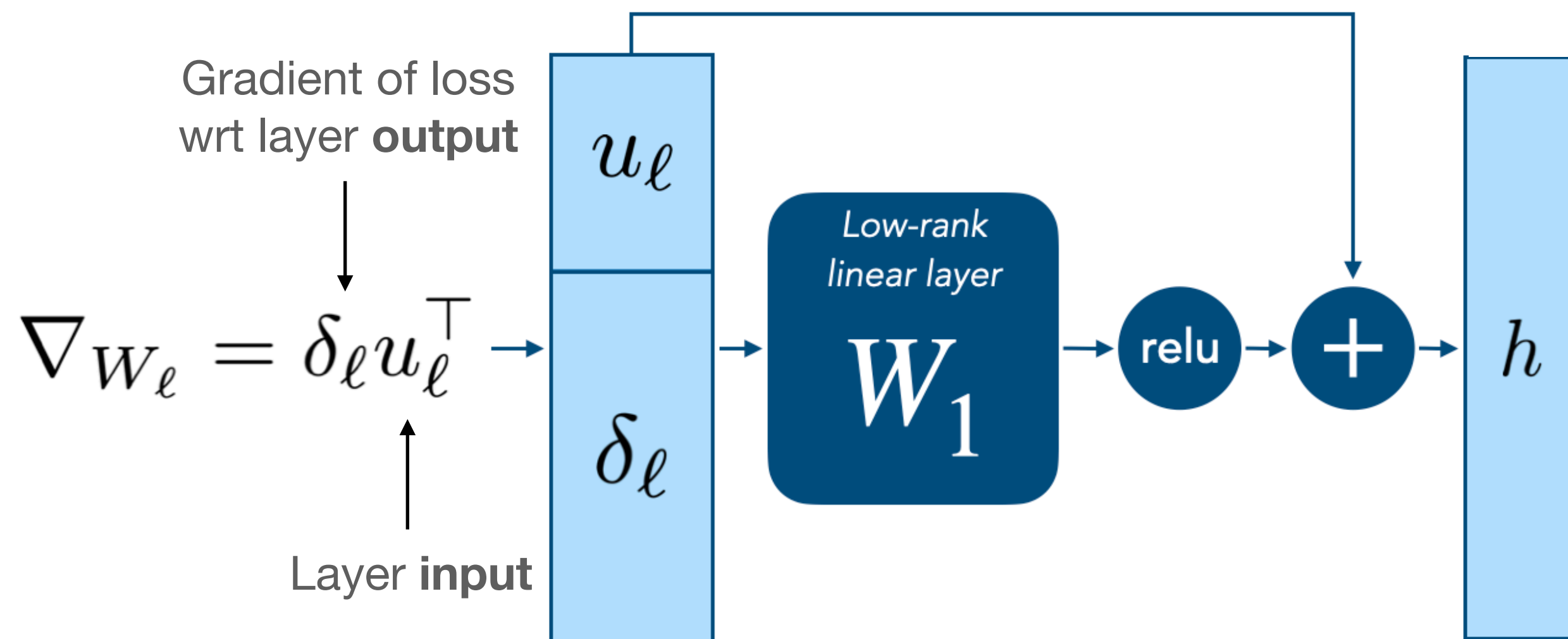
$$\nabla_{W_\ell} = \delta_\ell u_\ell^\top$$

↑

Layer **input**

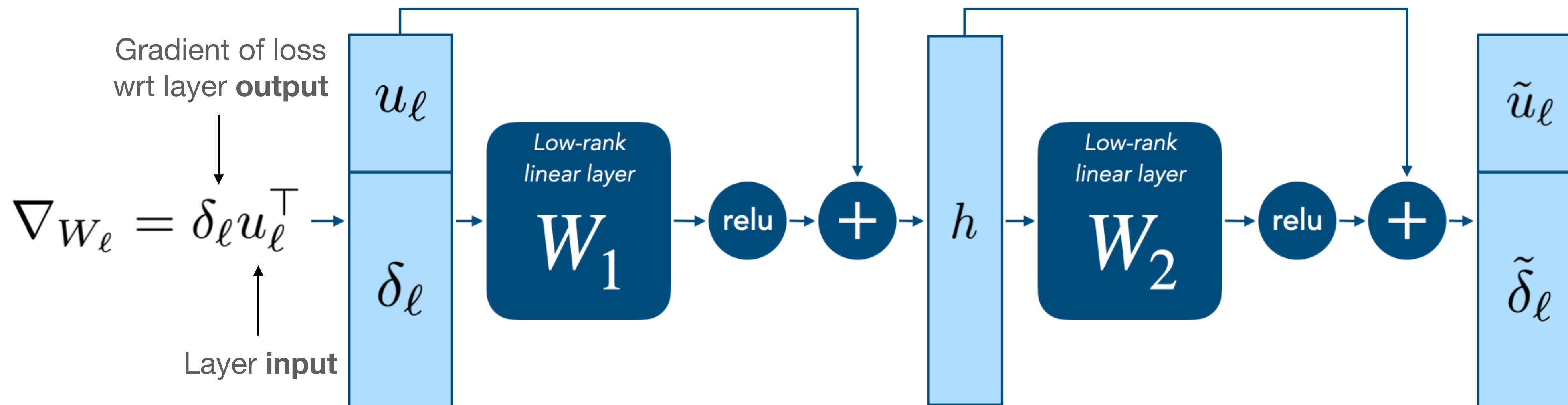
Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform



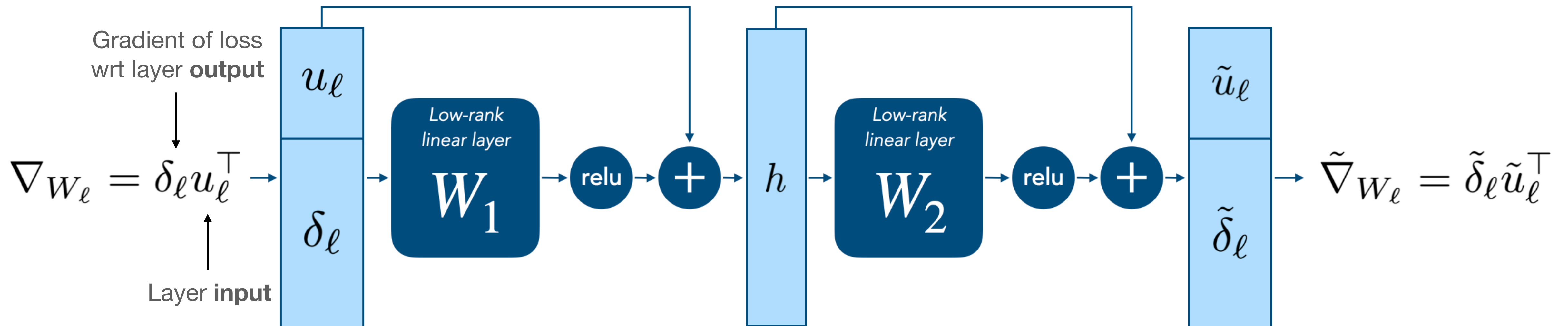
Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform



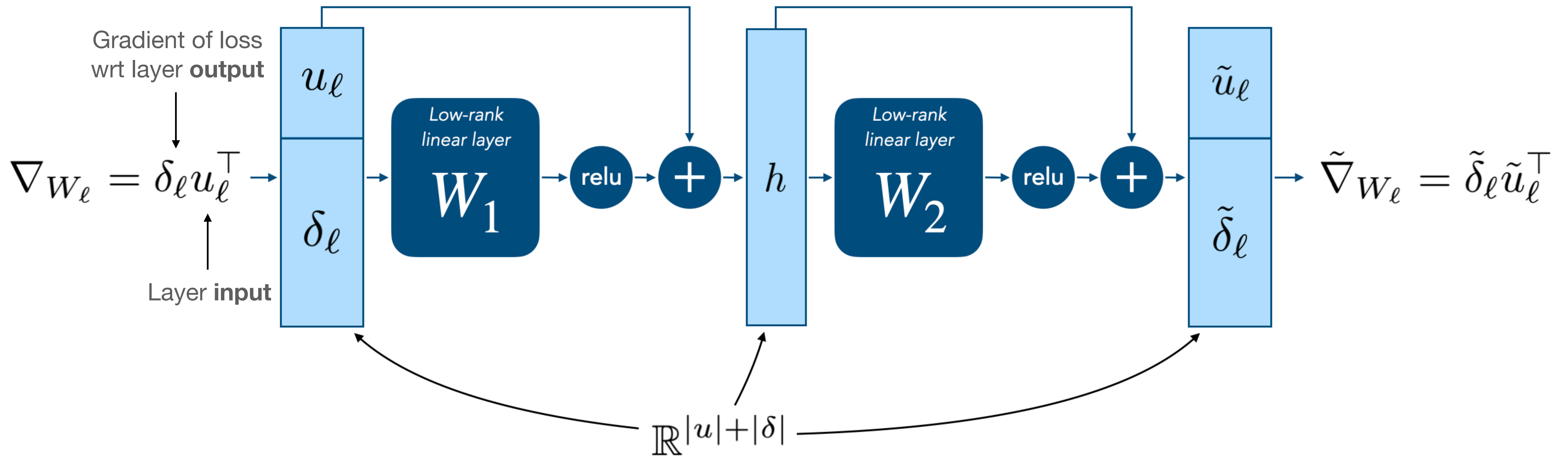
Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform



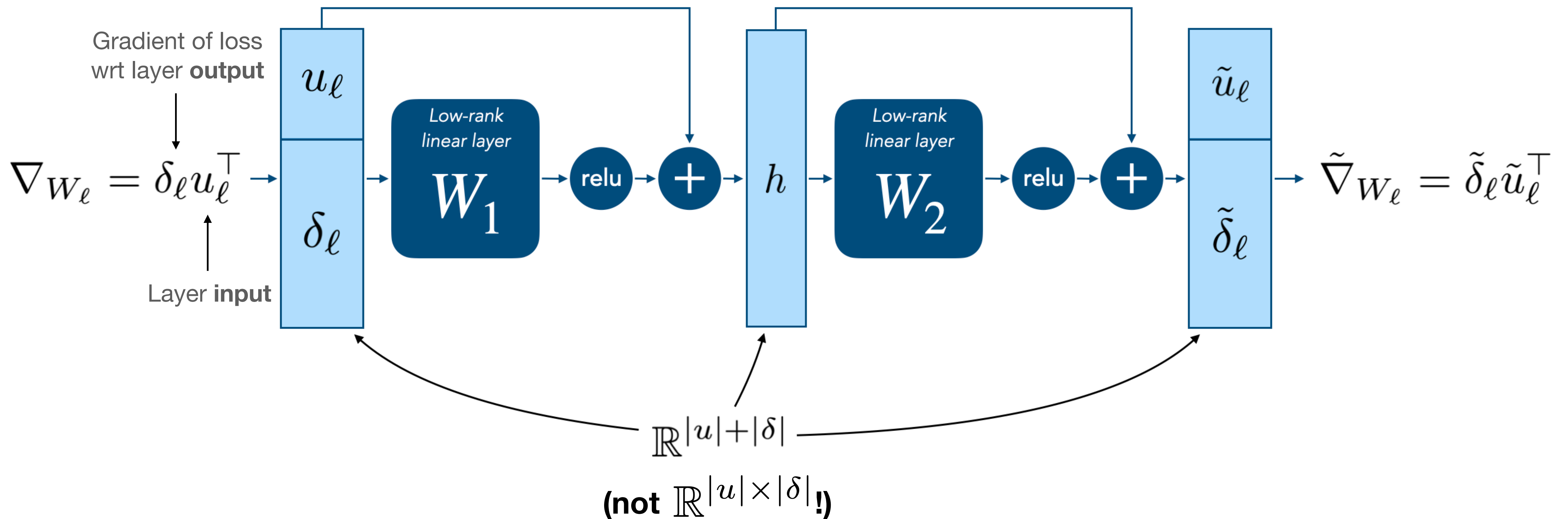
Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform



Model Editor Networks using Gradient Decomposition

An efficient, expressive gradient transform



Model Editor Networks using Gradient Decomposition

Editing T5-Large with multiple related edits at once

Input	Pre-Edit Output
Who is India's PM?	Satya Pal Malik ✗
Who is the prime minister of the UK?	Theresa May ✗
Who is the prime minister of India?	Narendra Modi ✓
Who is the UK PM?	Theresa May ✗

Bold text indicates the edits applied in each evaluation

Model Editor Networks using Gradient Decomposition

Editing T5-Large with multiple related edits at once

Input	Pre-Edit Output	Edit Target
Who is India's PM?	Satya Pal Malik ✗	Narendra Modi
Who is the prime minister of the UK?	Theresa May ✗	Boris Johnson
Who is the prime minister of India?	Narendra Modi ✓	-
Who is the UK PM?	Theresa May ✗	-

Bold text indicates the edits applied in each evaluation

Model Editor Networks using Gradient Decomposition

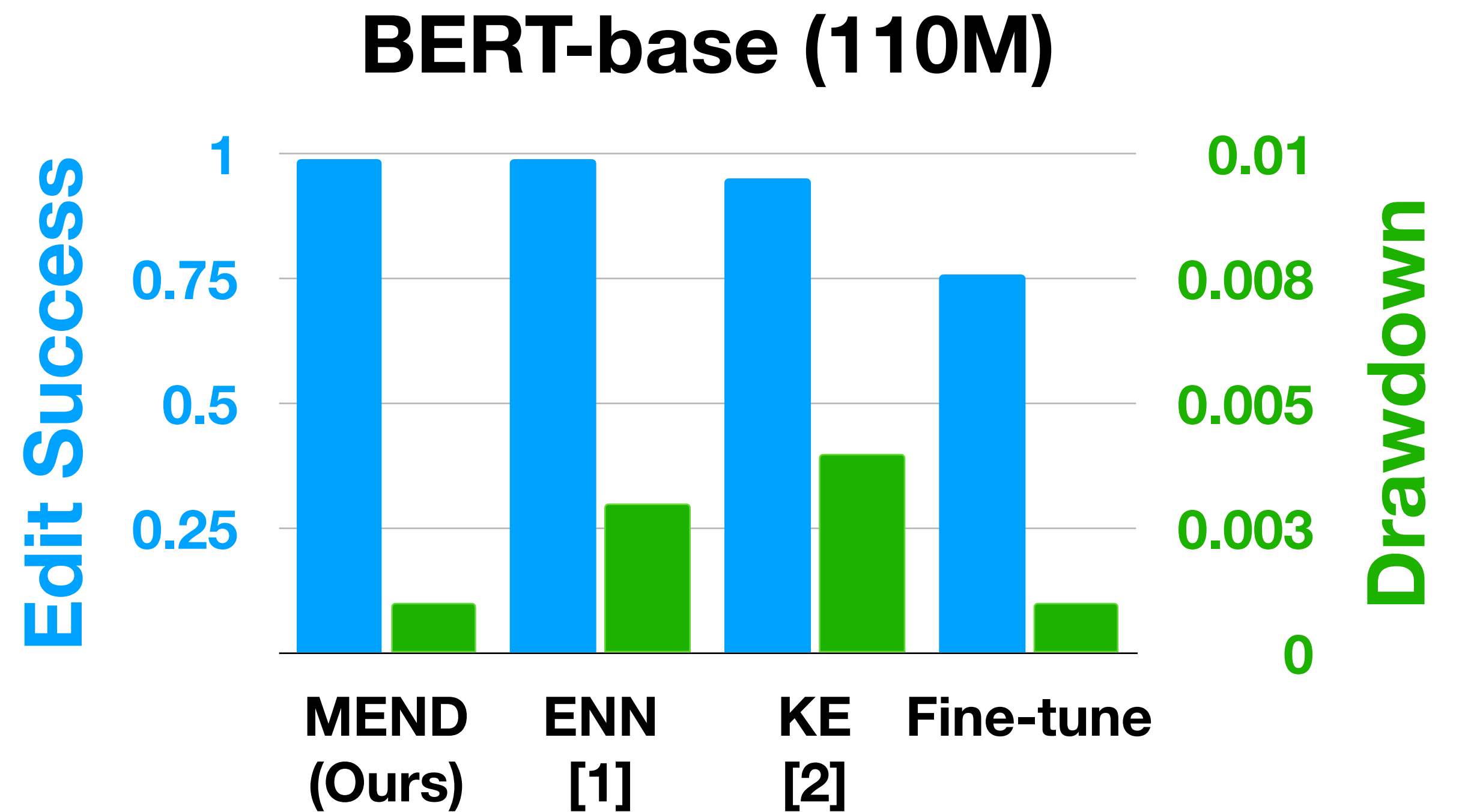
Editing T5-Large with multiple related edits at once

Input	Pre-Edit Output	Edit Target	Post-Edit Output
Who is India's PM?	Satya Pal Malik ✗	Narendra Modi	Narendra Modi ✓
Who is the prime minister of the UK?	Theresa May ✗	Boris Johnson	Boris Johnson ✓
Who is the prime minister of India?	Narendra Modi ✓	-	Narendra Modi ✓
Who is the UK PM?	Theresa May ✗	-	Boris Johnson ✓

Bold text indicates the edits applied in each evaluation

Model Editor Networks using Gradient Decomposition

Effective editing at small scale...

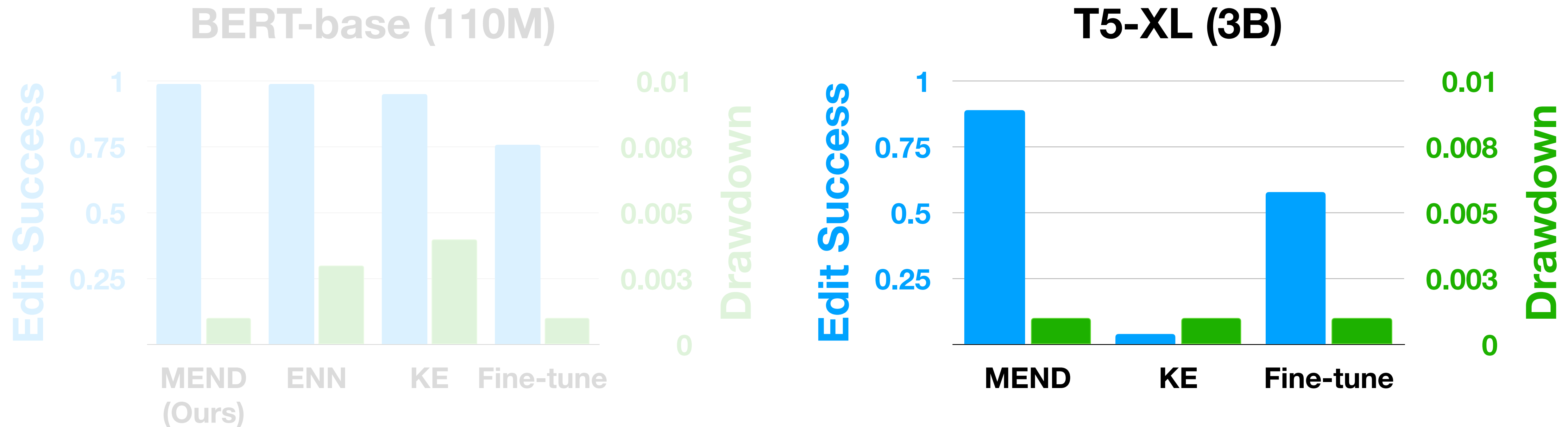


[1] Editable Neural Networks. Sinitsin et al. ICLR 2020.

[2] Editing Factual Knowledge in Language Models. De Cao et al. EMNLP 2021.

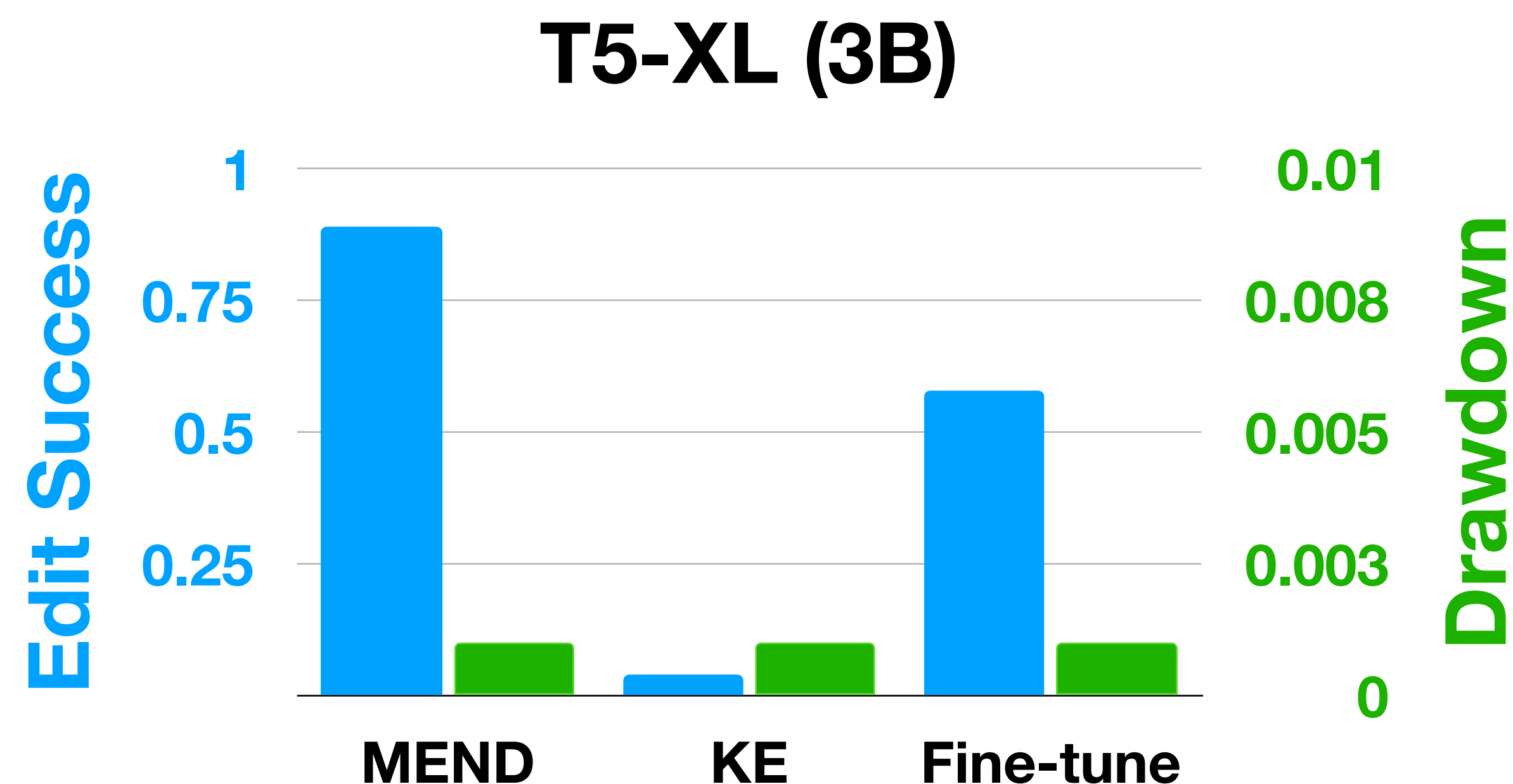
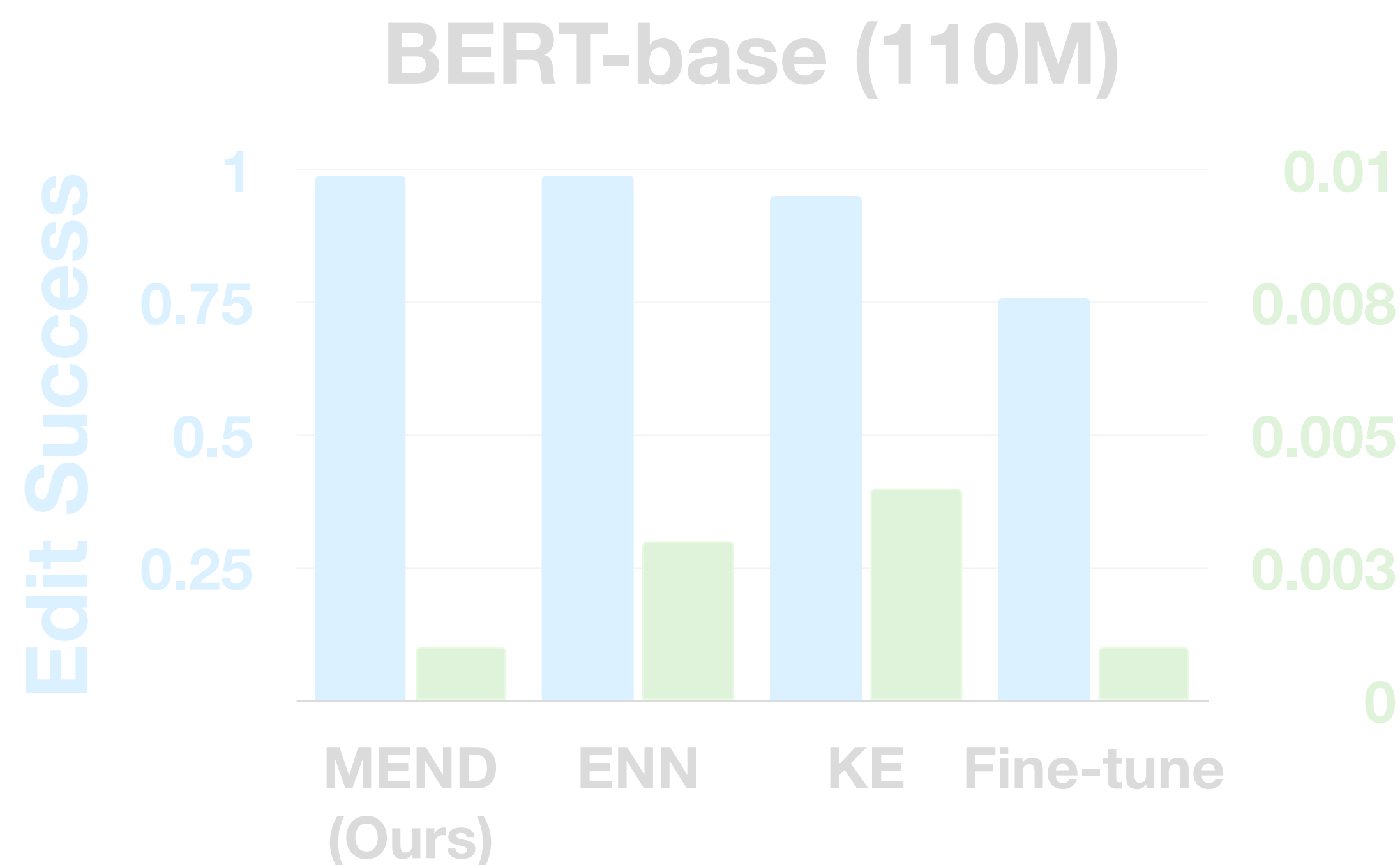
Model Editor Networks using Gradient Decomposition

Effective editing at small scale...and large scale!



Model Editor Networks using Gradient Decomposition

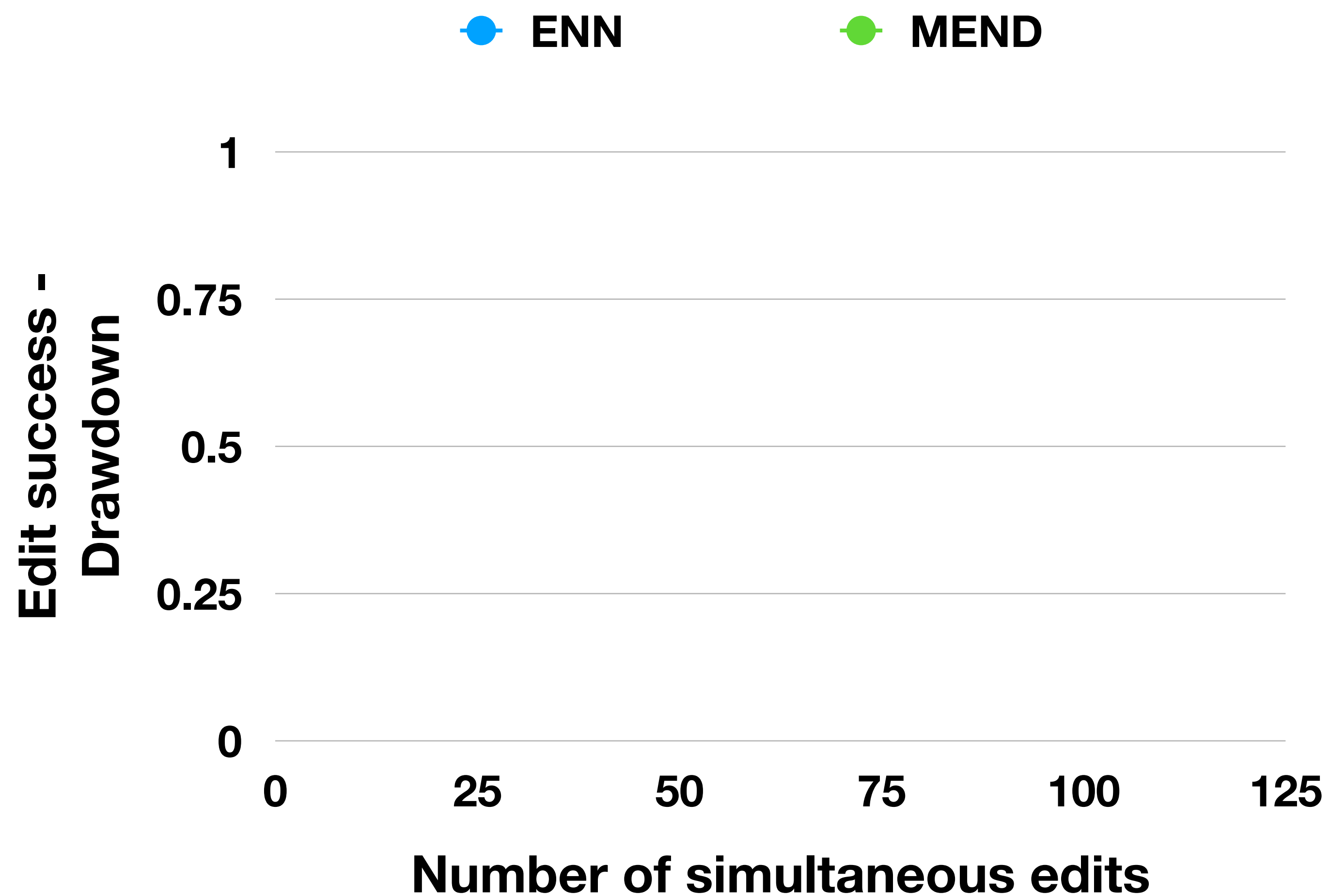
Effective editing at small scale...and large scale!



*ENN gives OOM

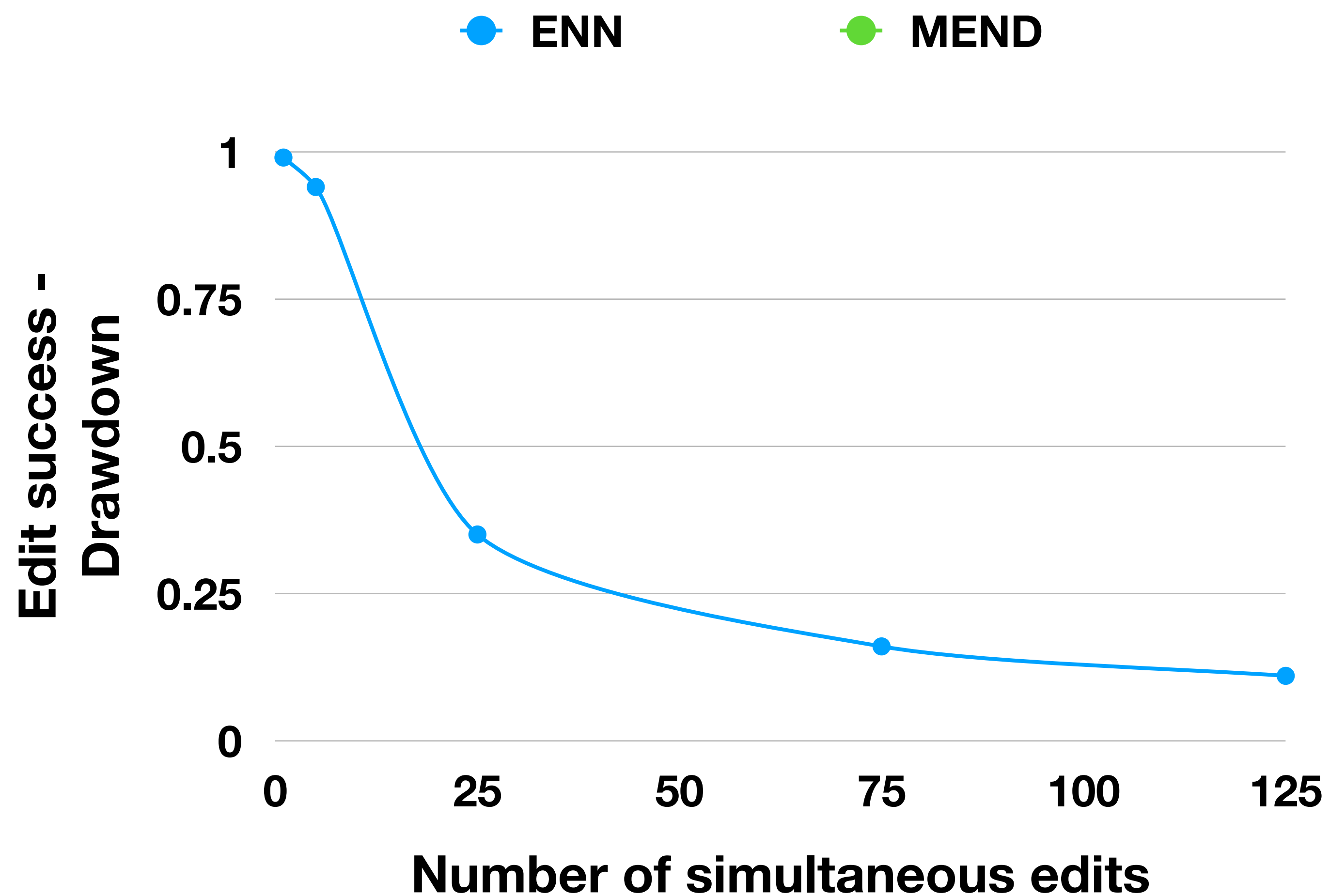
Inching towards the real world...

Applying multiple edits to BART-base



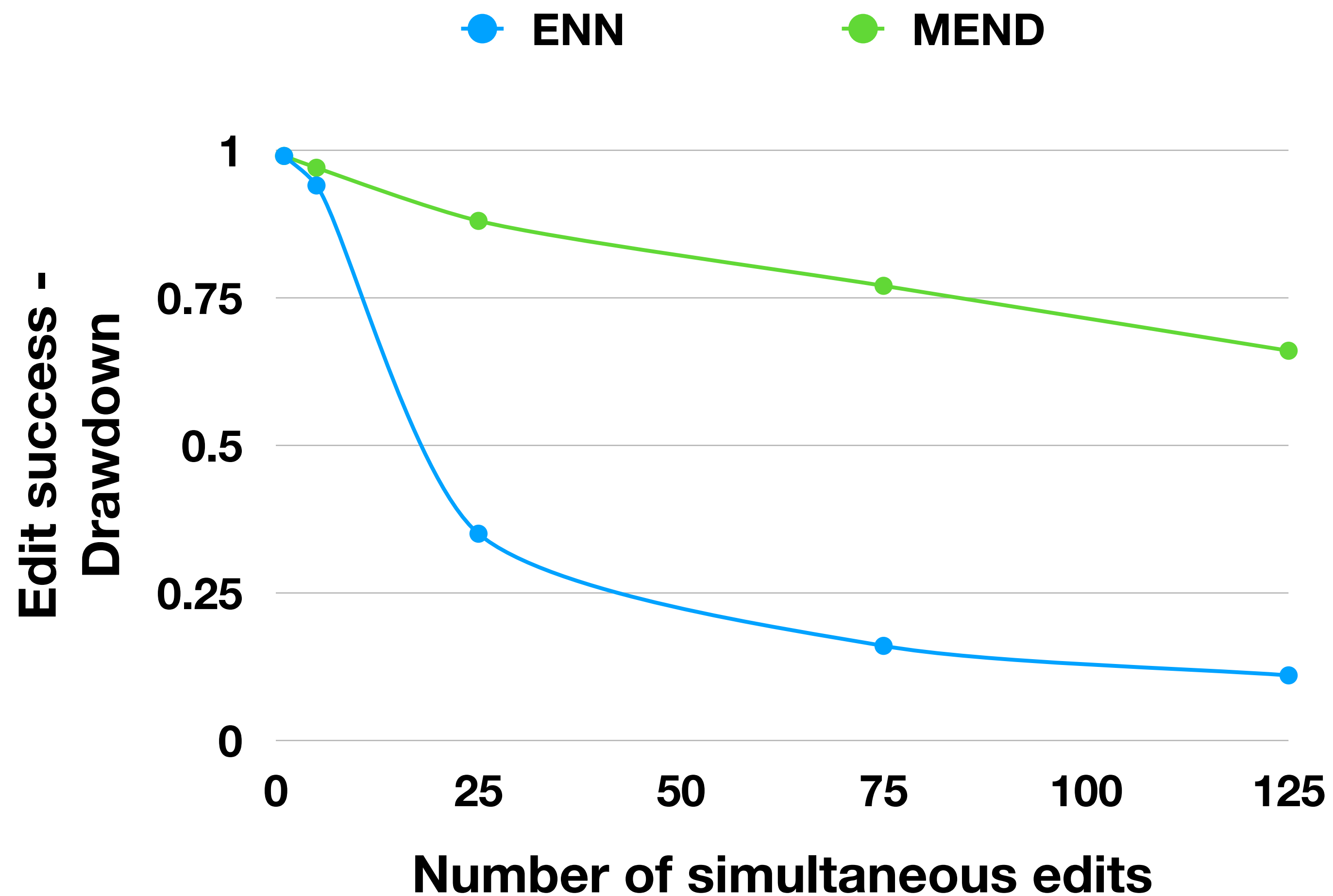
Inching towards the real world...

Applying multiple edits to BART-base



Inching towards the real world...

Applying multiple edits to BART-base



Conclusion

- Large models become widespread → model errors **impact more people**

Conclusion

- Large models become widespread → model errors **impact more people**
- **Model editors** can enable cheaper/faster harm mitigation & increase uptime

Conclusion

- Large models become widespread → model errors **impact more people**
- **Model editors** can enable cheaper/faster harm mitigation & increase uptime
- MEND enables fast, effective editing on models with **10B+ parameters**

Conclusion

- Large models become widespread → model errors **impact more people**
- **Model editors** can enable cheaper/faster harm mitigation & increase uptime
- MEND enables fast, effective editing on models with **10B+ parameters**

Paper: `tinyurl.com/mend-iclr`

Code: `github.com/eric-mitchell/mend`



Paper & code