

Memory-Based Model Editing at Scale

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

Stanford University

2022 International Conference on Machine Learning



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!

Who is the president of the US? Joe Biden

Who is the prime minister of the UK? Theresa May

Who is the president of Russia? Vladimir Putin

Who is the president of China? Xi Jinping

Who is the president of France? Emmanuel Macron

Who is the president of Germany? Angela Merkel

Who is the president of Nigeria? Muhammadu Buhari

Who is the president of the US? Donald Trump

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

Who is the president of the US? Joe Biden

Who is the prime minister of the UK? Theresa May

Who is the president of Russia? Vladimir Putin

Who is the president of China? Xi Jinping

Who is the president of France? Emmanuel Macron

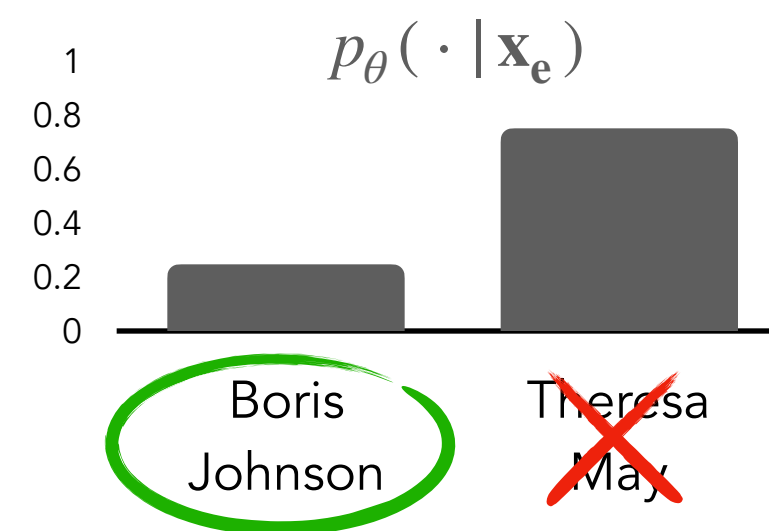
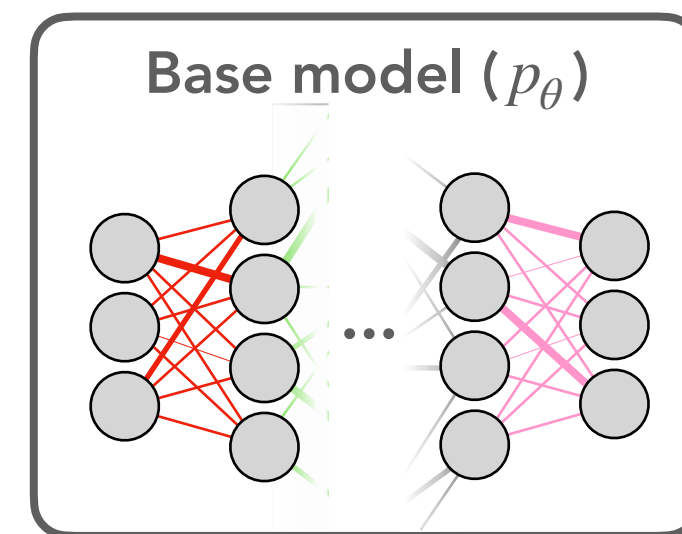
Who is the president of Germany? Angela Merkel

Who is the president of Nigeria? Muhammadu Buhari

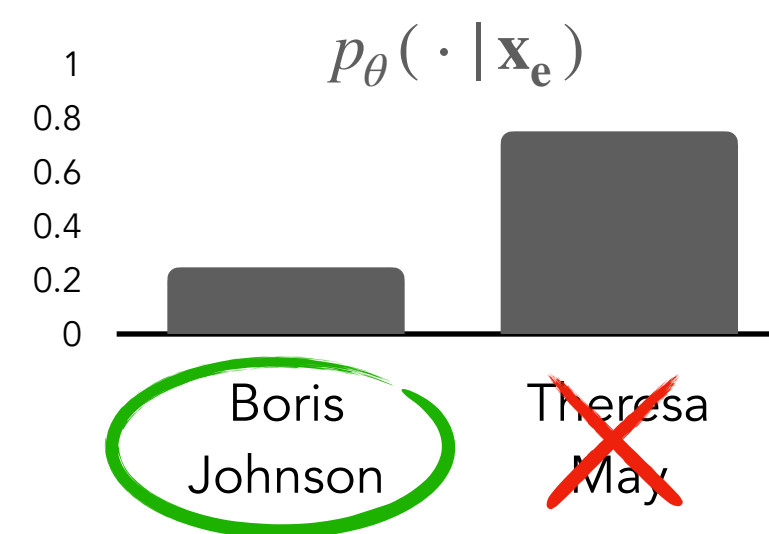
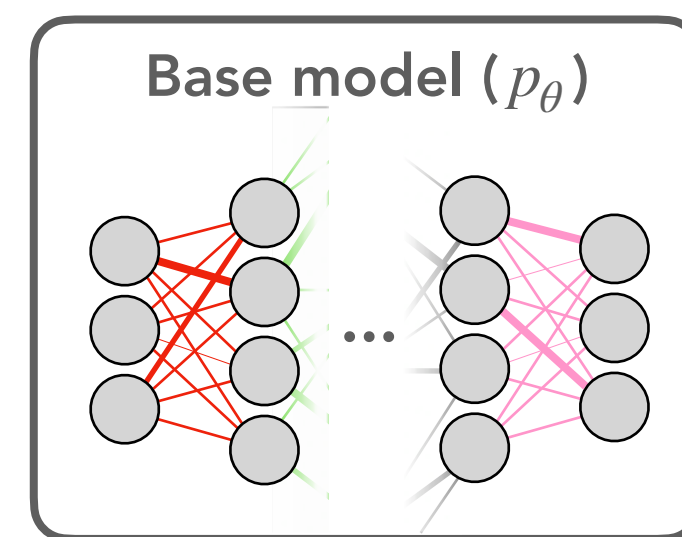
Who is the president of the US? Donald Trump

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

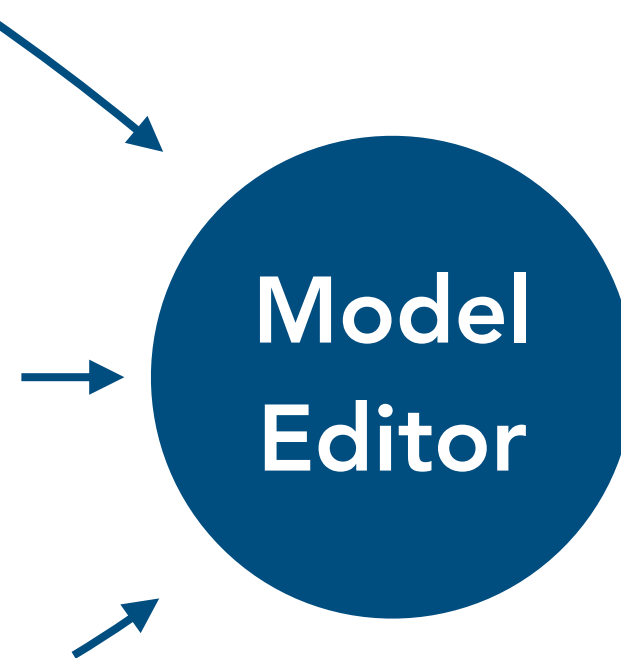
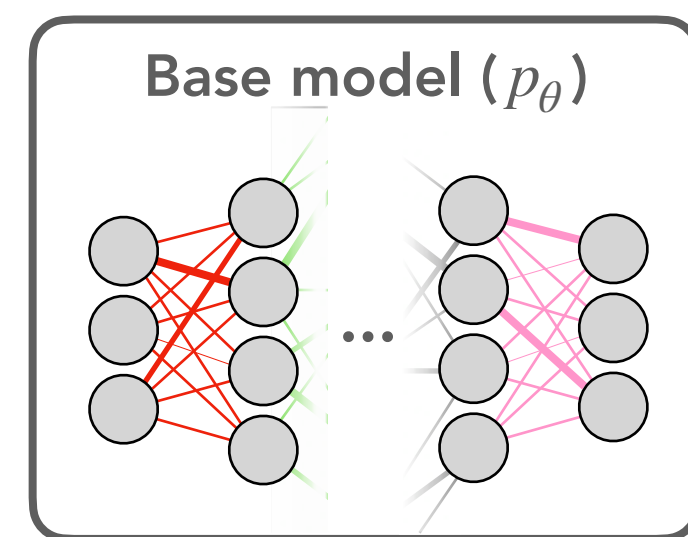
$\mathbf{x}_e =$ "Who is the
prime minister
of the UK?"



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



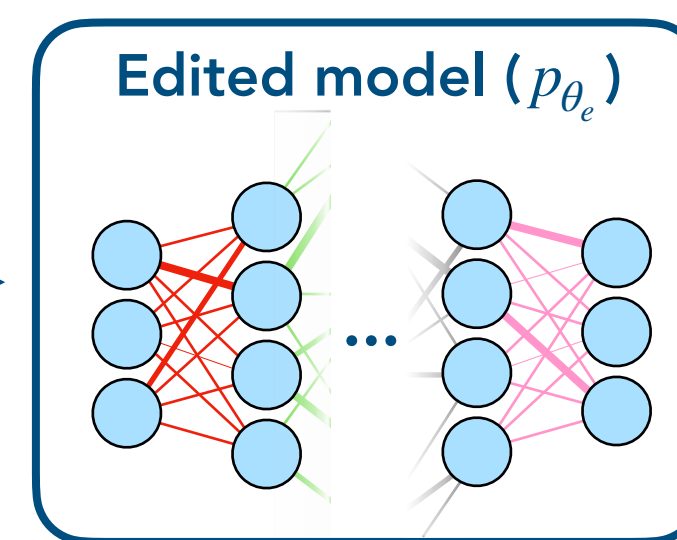
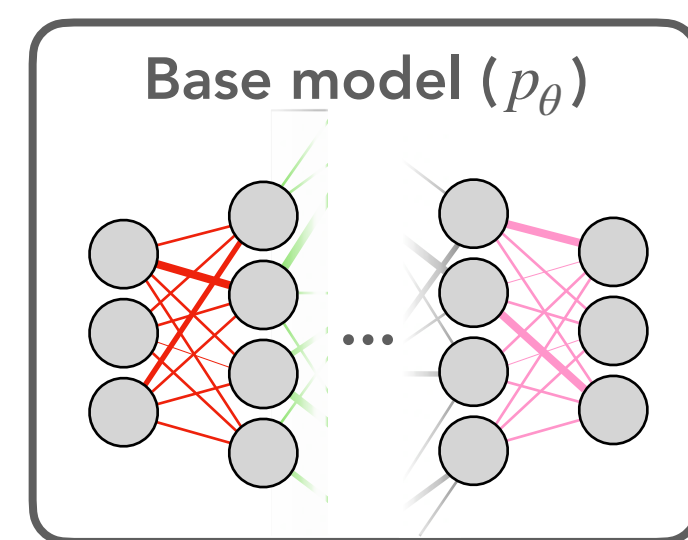
$\mathbf{x}_e =$ "Who is the
prime minister
of the UK?"



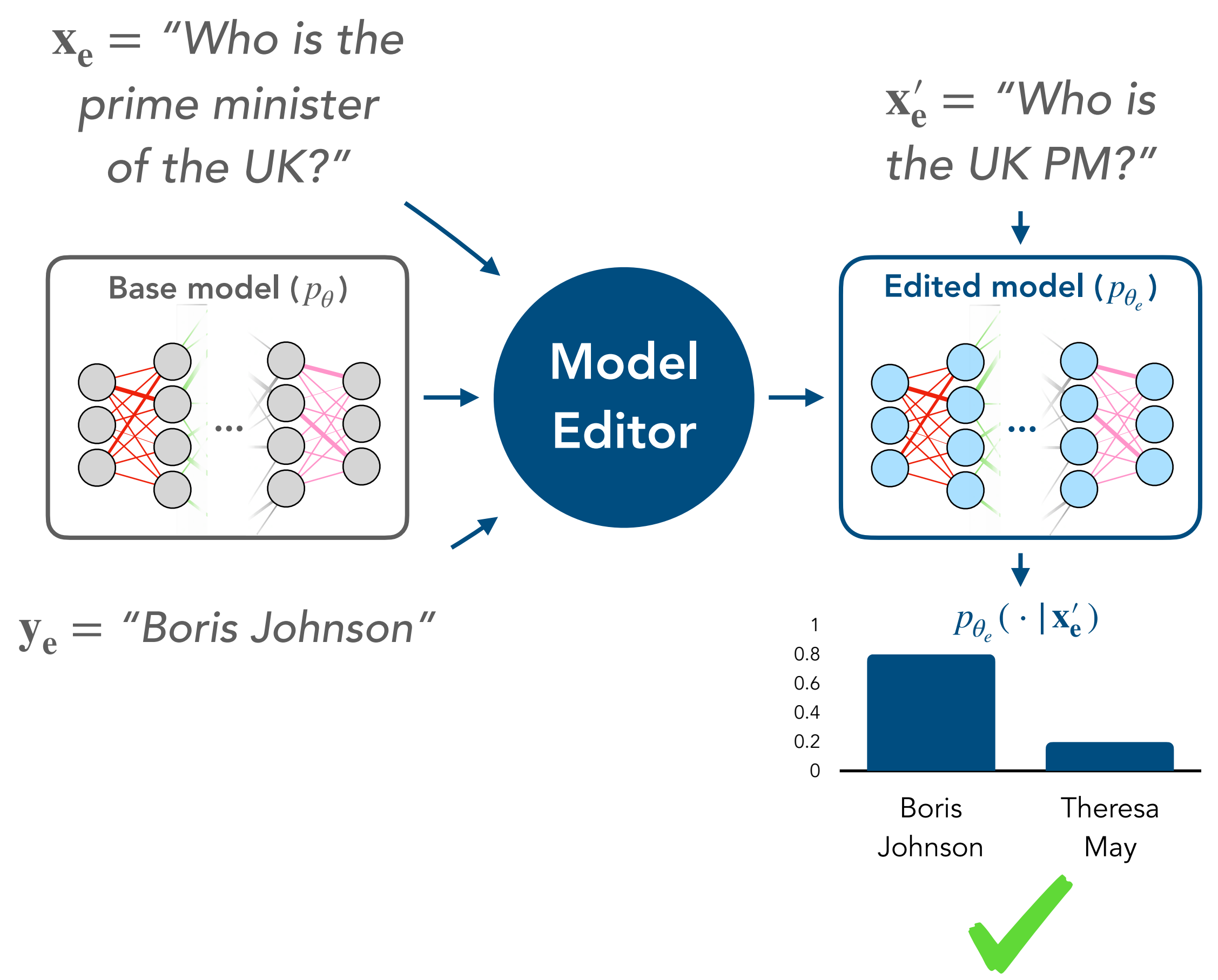
Model
Editor

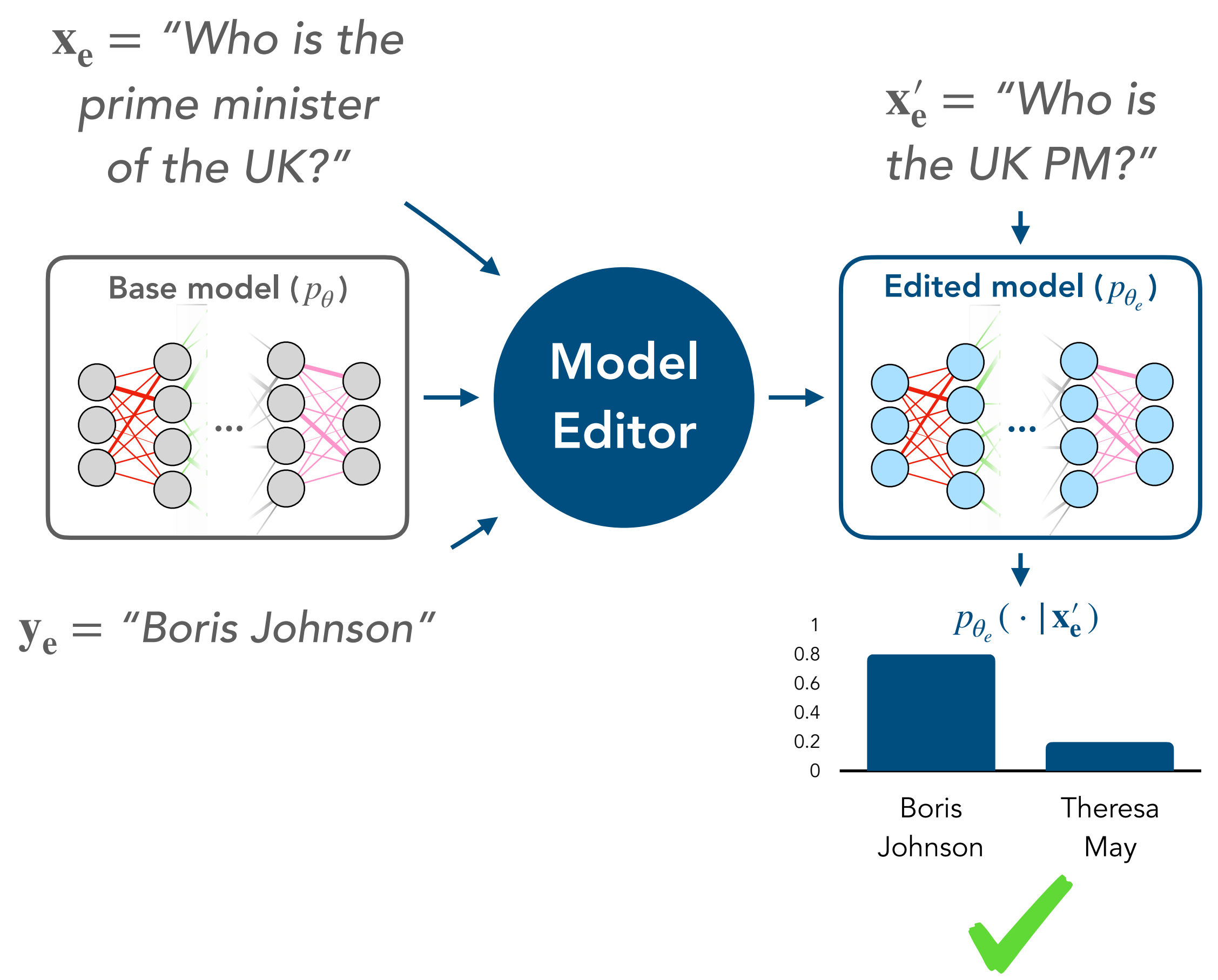
$\mathbf{y}_e =$ "Boris Johnson"

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

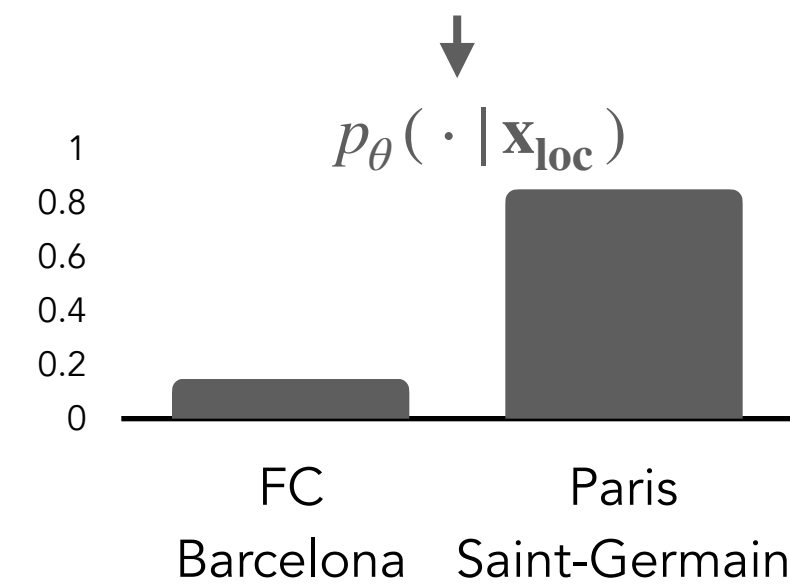
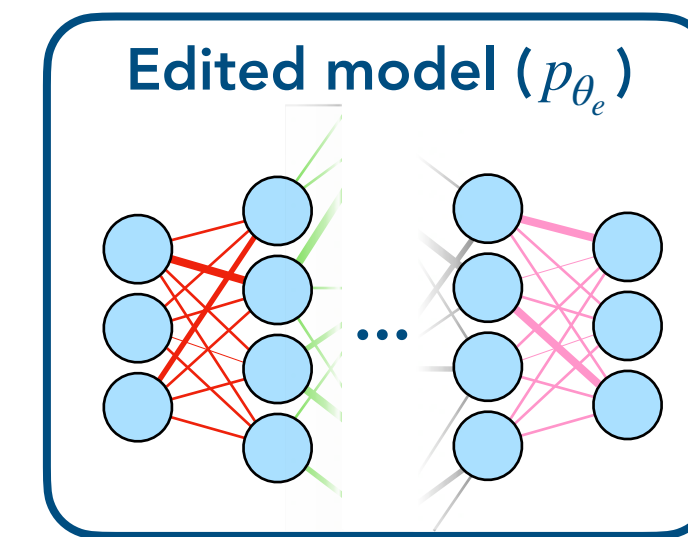
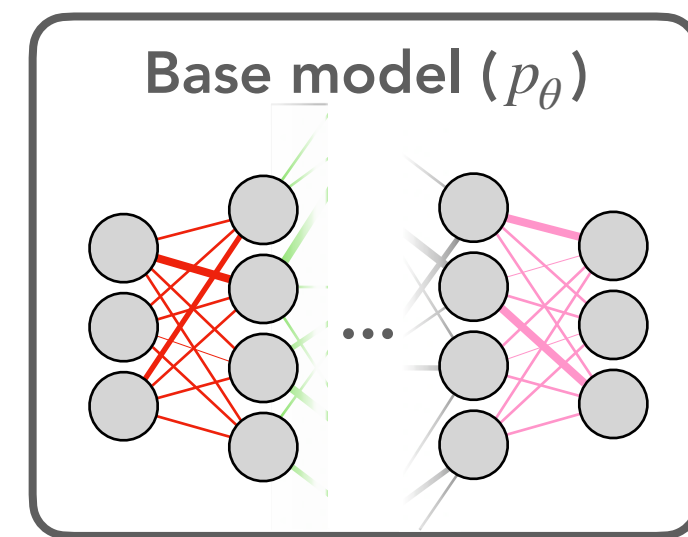


$\mathbf{y}_e =$ "Boris Johnson"

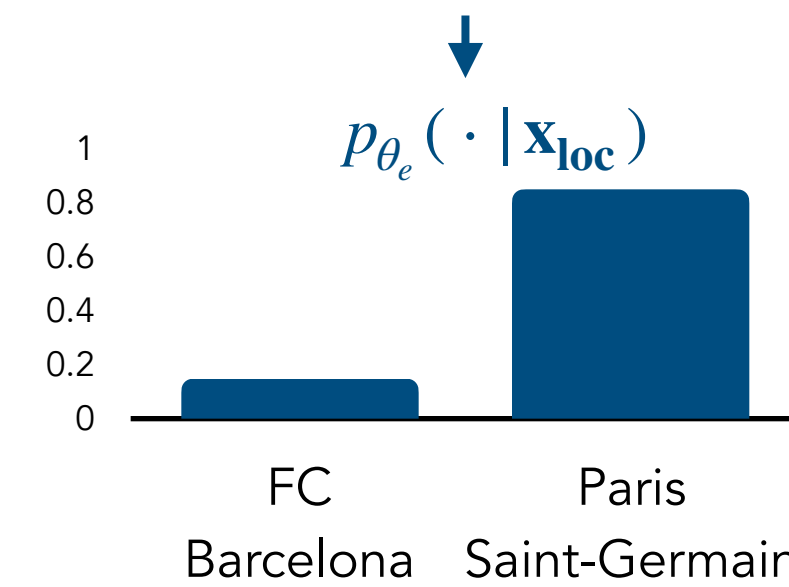




$\mathbf{x}_{\text{loc}} = \text{"Who does Messi play for?"}$



Unchanged
by edit



Edit *what*, exactly?

Defining the problem

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

Edit example



Edit *what*, exactly?

Defining the problem



Edit example

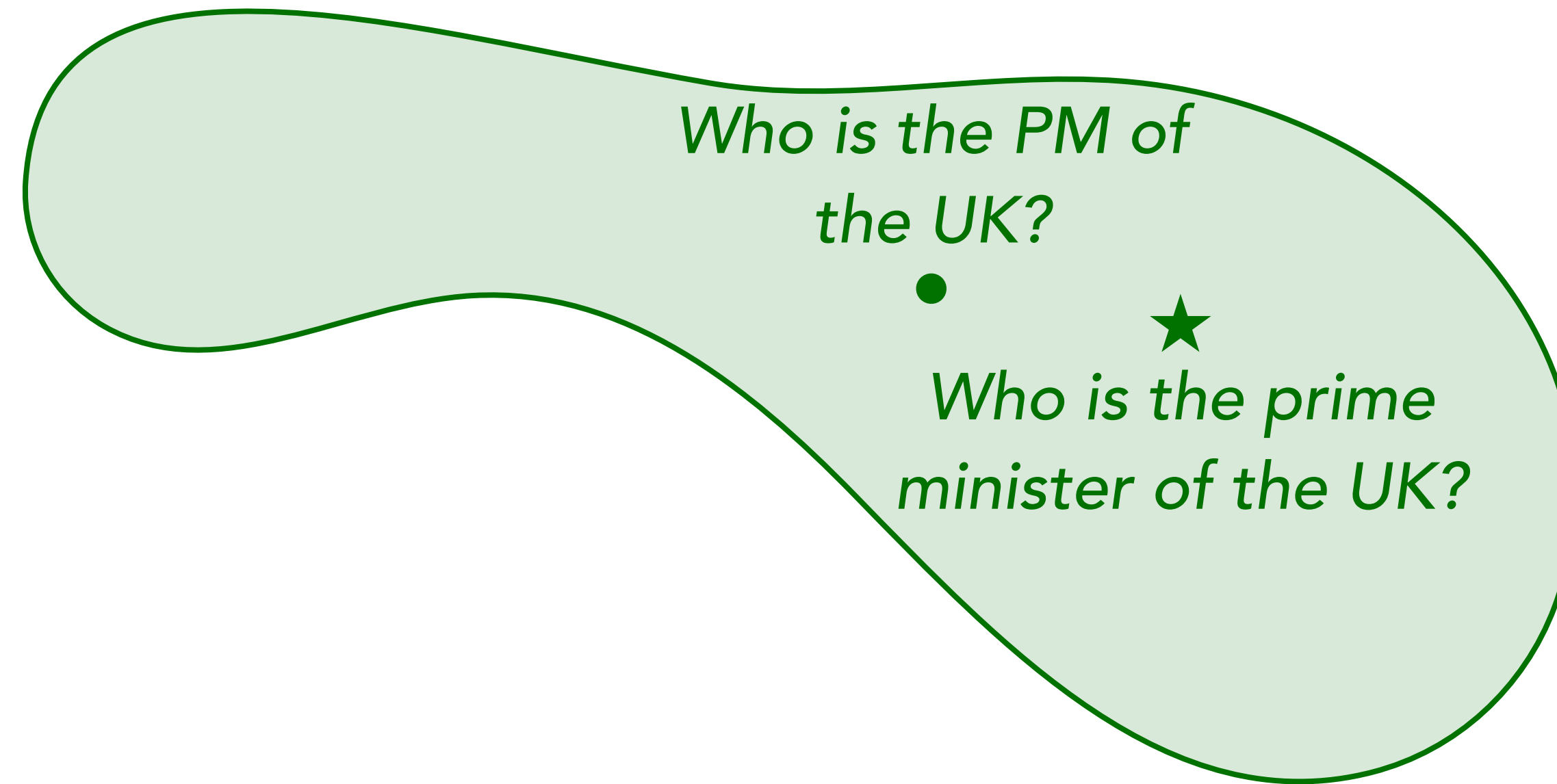


Edit scope



Edit *what*, exactly?

Defining the problem



Edit example



Edit scope

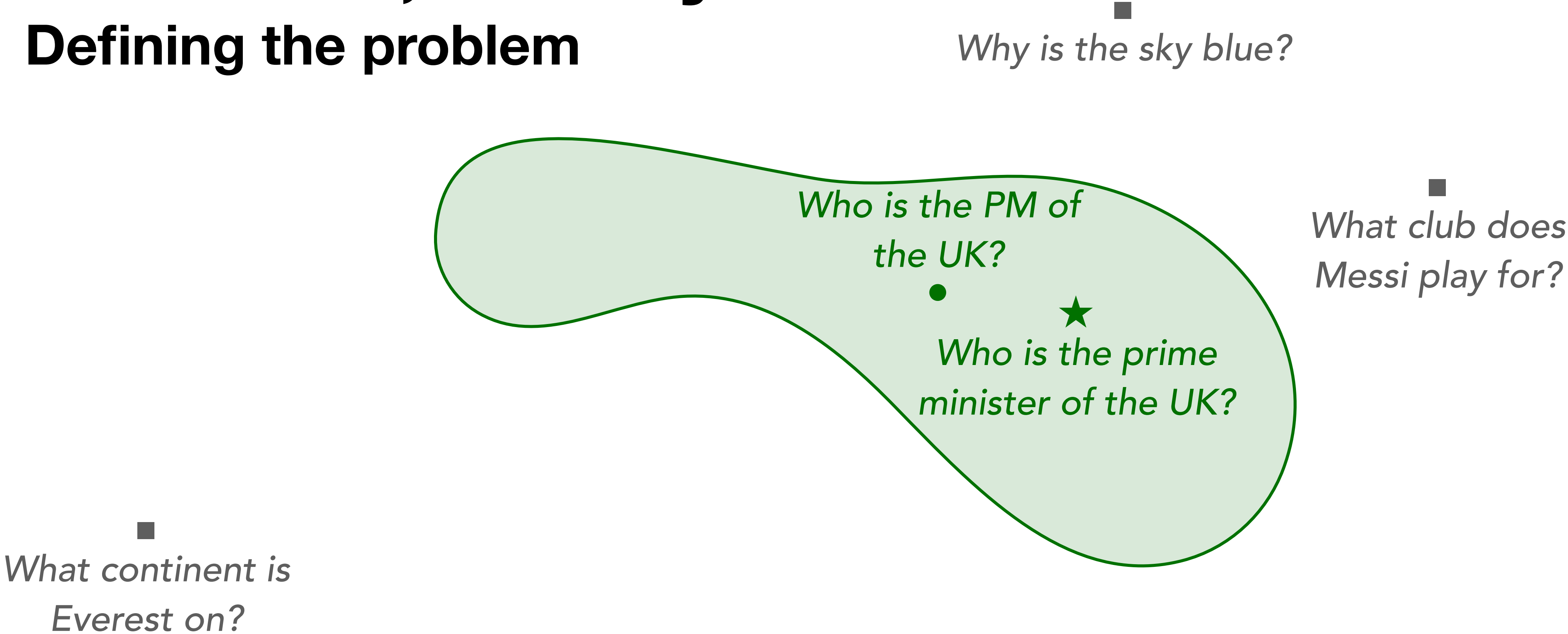


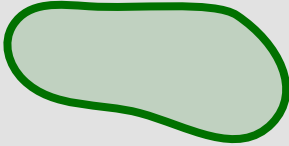
In-scope



Edit *what*, exactly?

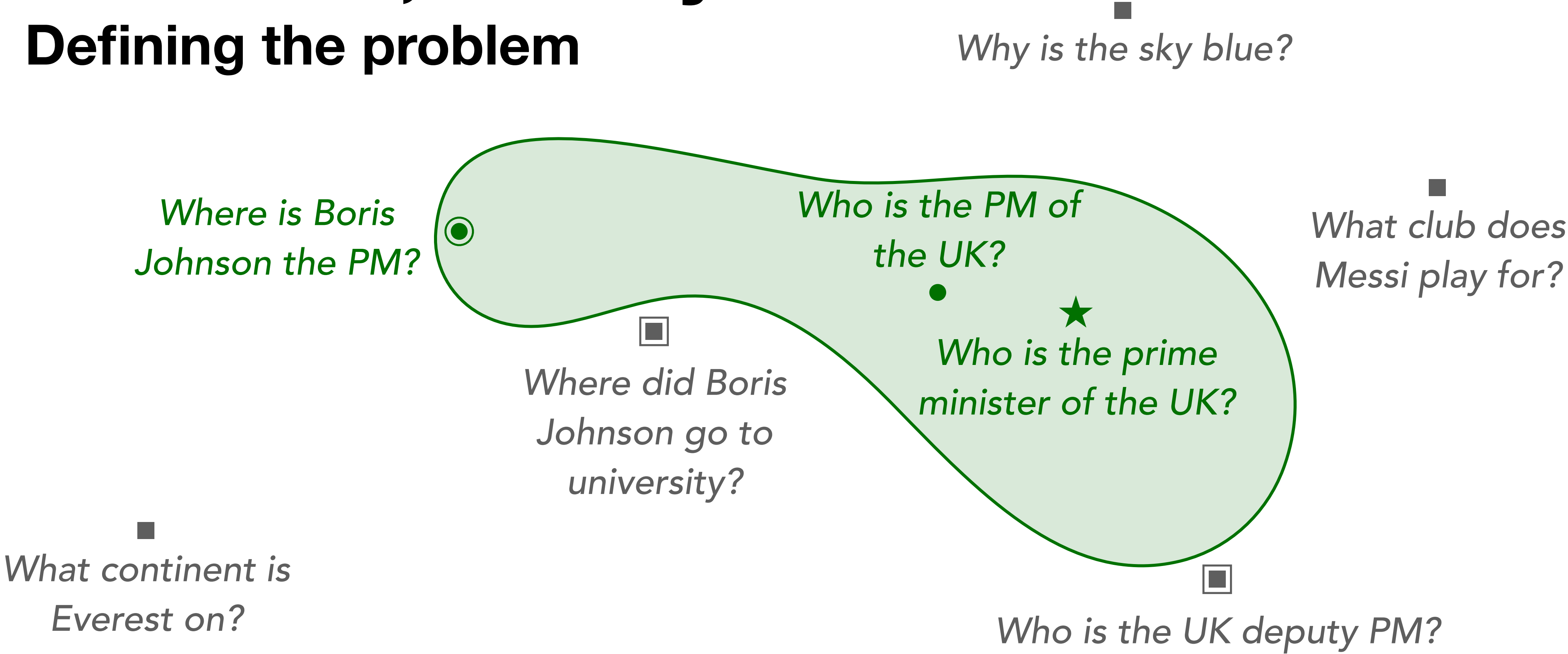
Defining the problem




Edit example	Edit scope	In-scope	Out-of-scope
★		●	■

Edit *what*, exactly?

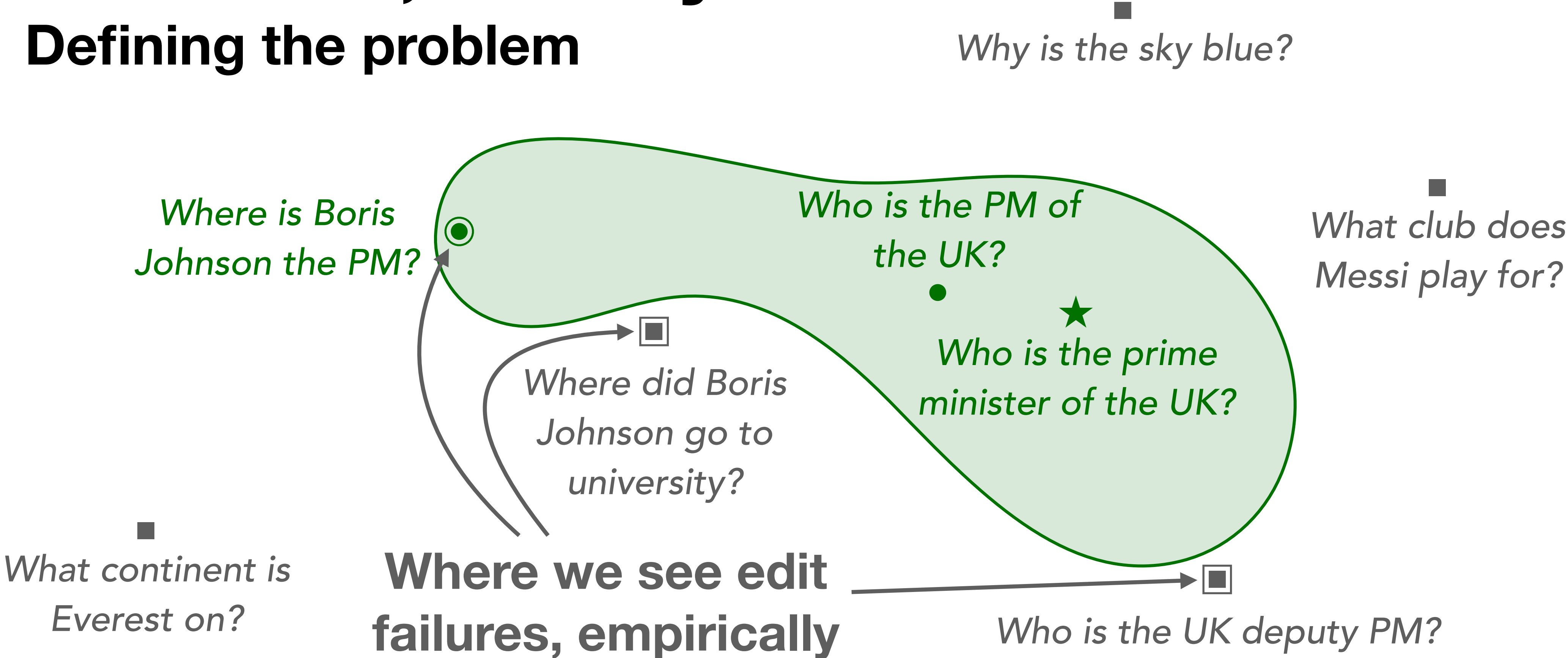
Defining the problem




Edit example	Edit scope	In-scope	Out-of-scope	Hard in/out-of-scope
★		●	■	⊙ □

Edit *what*, exactly?

Defining the problem

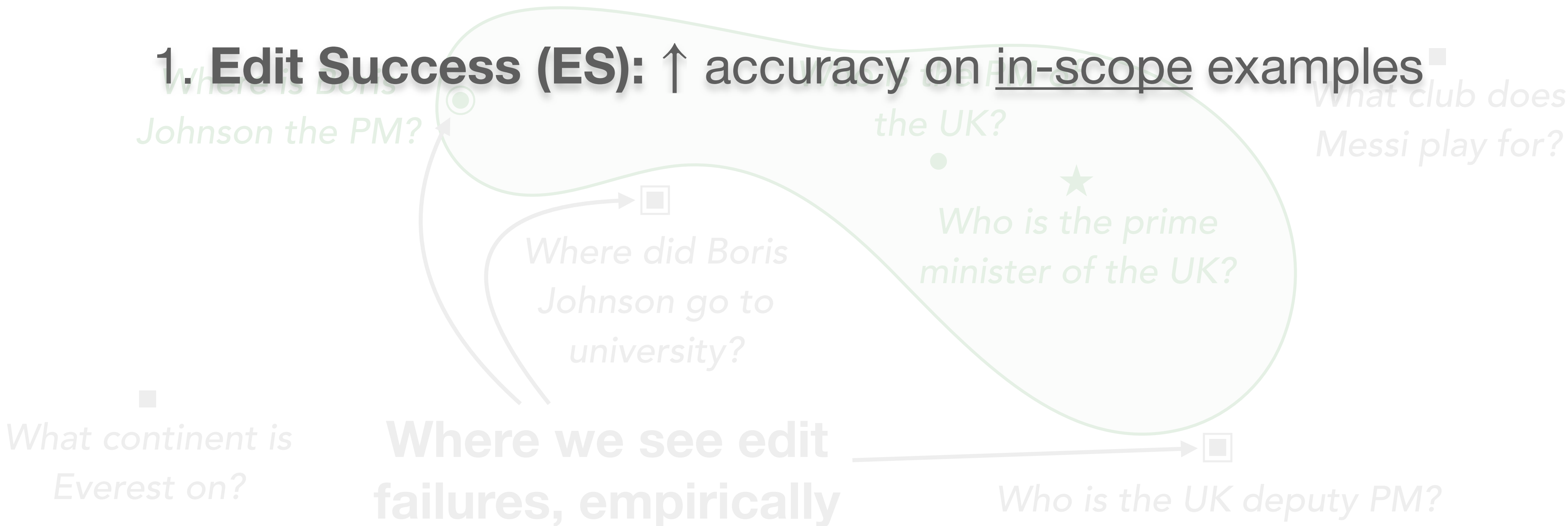







Edit example	Edit scope	In-scope	Out-of-scope	Hard in/out-of-scope
★		●	■	◎ □

Edit *what*, exactly?

Metrics for evaluating model edits

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

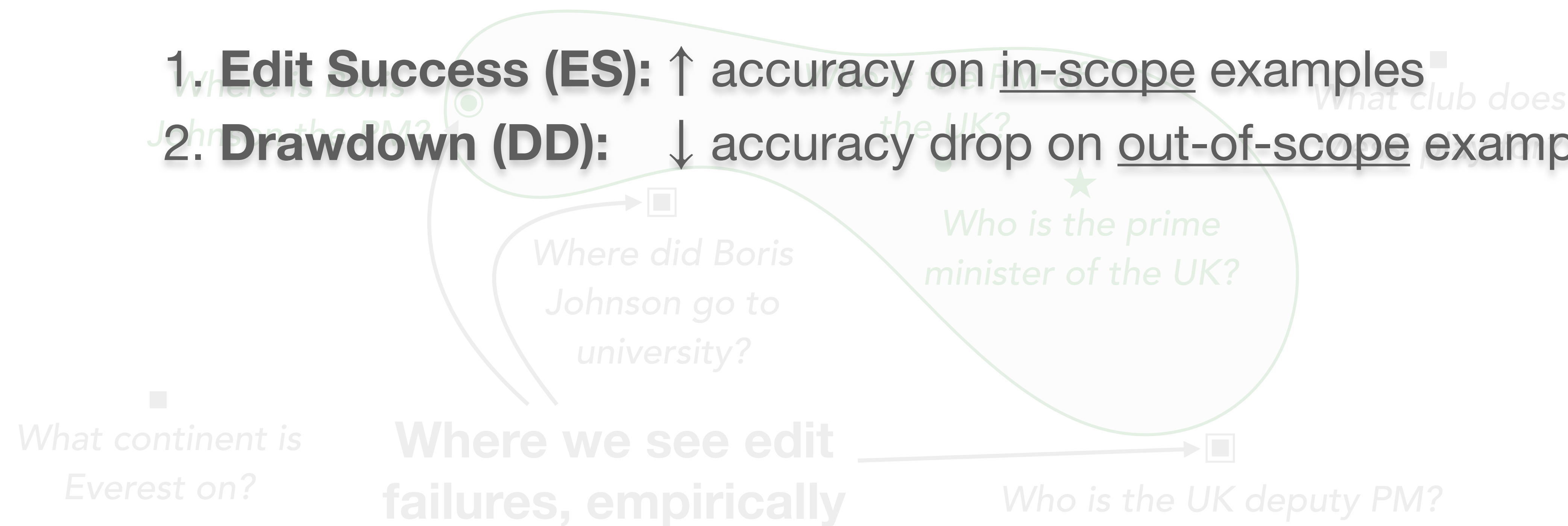


Edit scope	Edit example	In-scope	Out-of-scope	Hard in/out-of-scope
				

Edit *what*, exactly?

Metrics for evaluating model edits

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



Edit scope	Edit example	In-scope	Out-of-scope	Hard in/out-of-scope

Learning to edit

Editing as meta-learning

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

Learning to edit

Editing as meta-learning

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

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

$\mathbf{x}_{\text{loc}} =$ “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”

Learning to edit

Editing as meta-learning

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

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

$\mathbf{x}_{\text{loc}} =$ “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”

Learning to edit

Editing as meta-learning

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

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

$\mathbf{x}_{\text{loc}} =$ “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”

Enforce locality with
out-of-scope example

Learning to edit

Editing as meta-learning

Requirement: an “edit dataset” $\mathbf{D}_{\text{edit}} = \{ (\underbrace{\mathbf{z}_{\text{edit}}}_{\text{Perform edit}}, \underbrace{\mathbf{x}_{\text{loc}}, \mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}}}_{\text{Enforce generalization with in-scope example}}) \}$

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

$\mathbf{x}_{\text{loc}} =$ “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”

Enforce locality with
out-of-scope example

Enforce generalization
with in-scope example

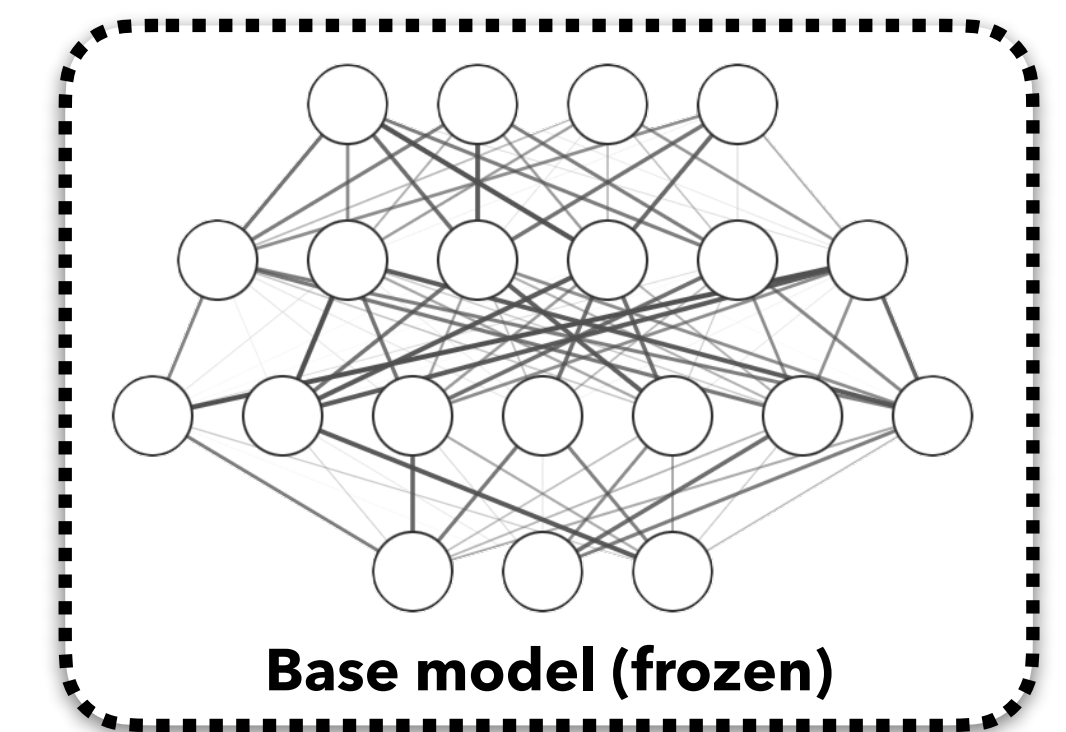
Edits without parameter updates

Semi-parametric Eediting with a Retrieval-Augmented Counterfactual model

Edits without parameter updates

Semi-parametric Eediting with a Retrieval-Augmented Counterfactual model

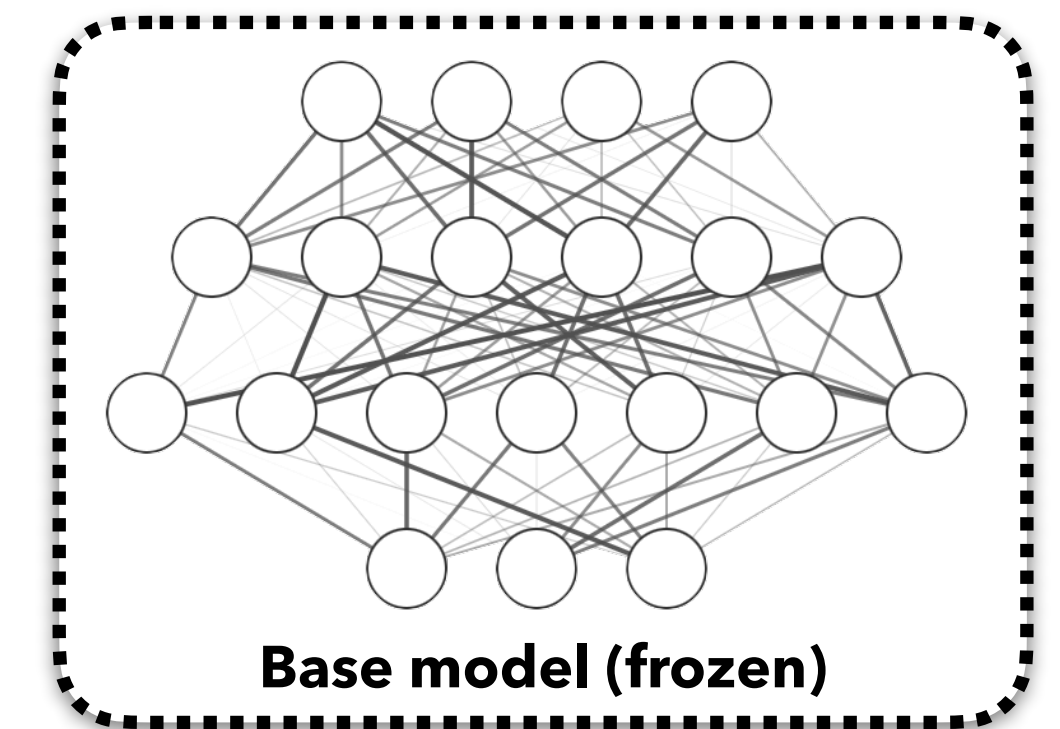
Start with the **frozen** base model



Edits without parameter updates

Semi-parametric Eediting with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

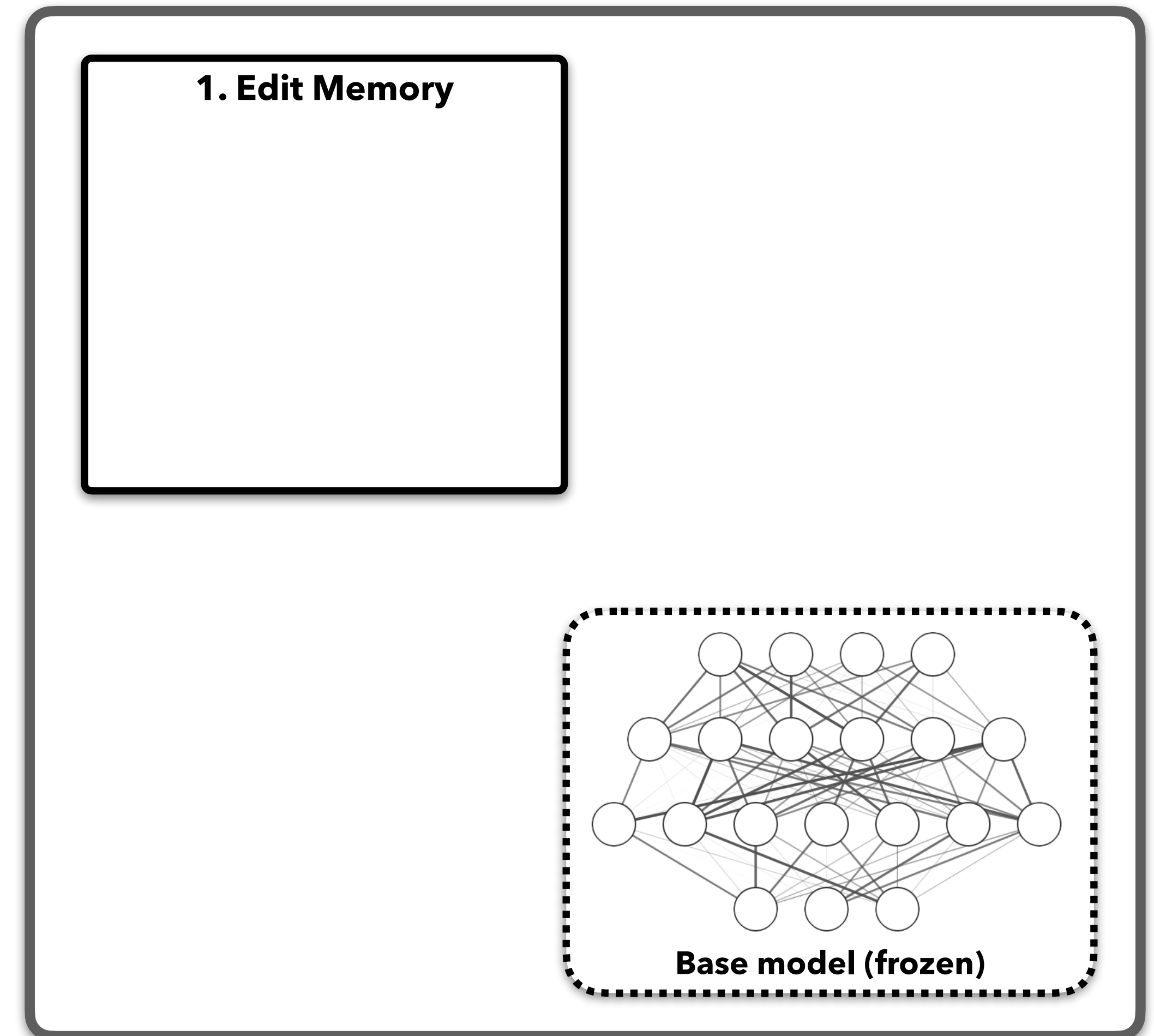


Edits without parameter updates

Semi-parametric Eediting with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

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

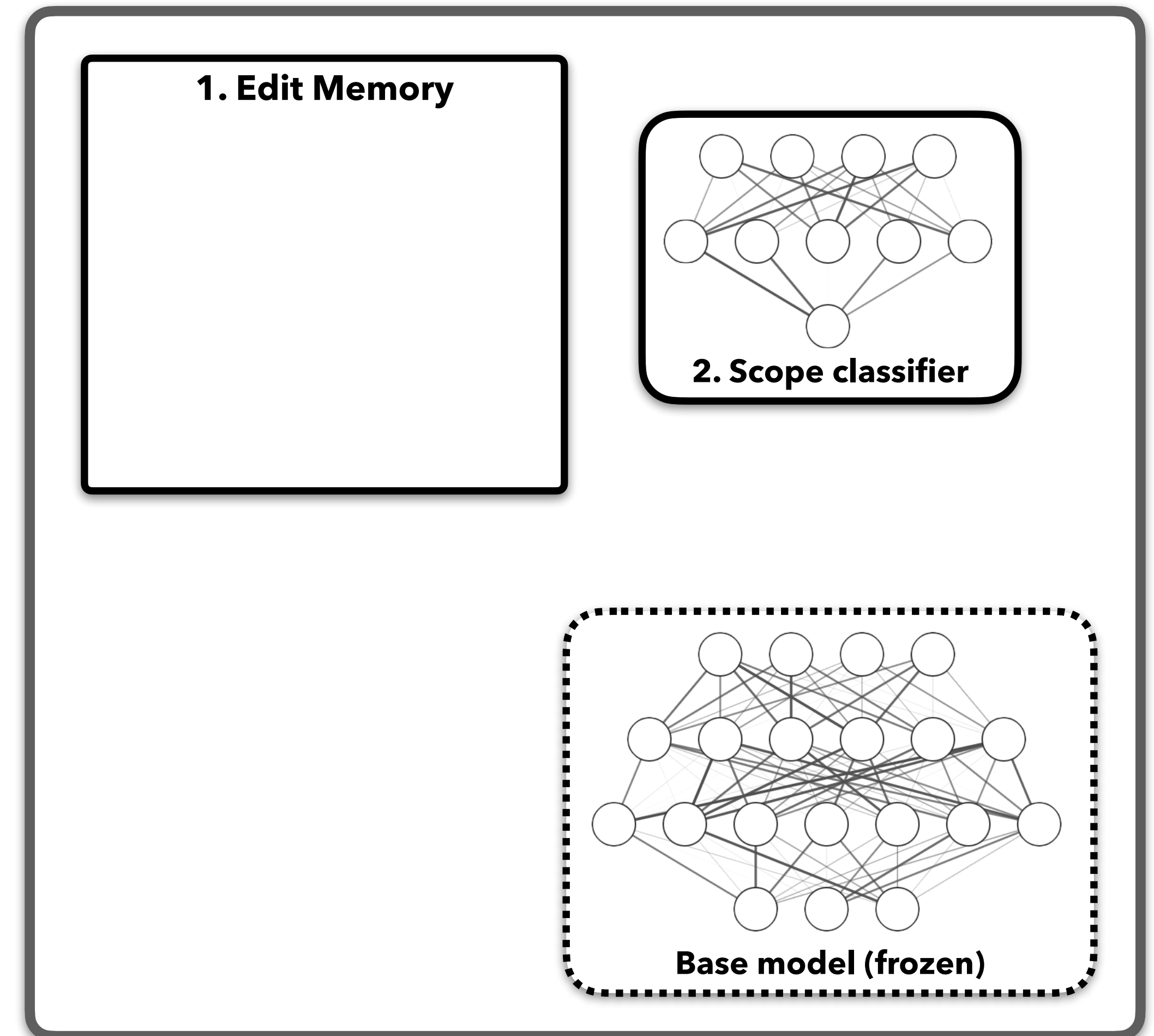


Edits without parameter updates

Semi-parametric Eediting 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

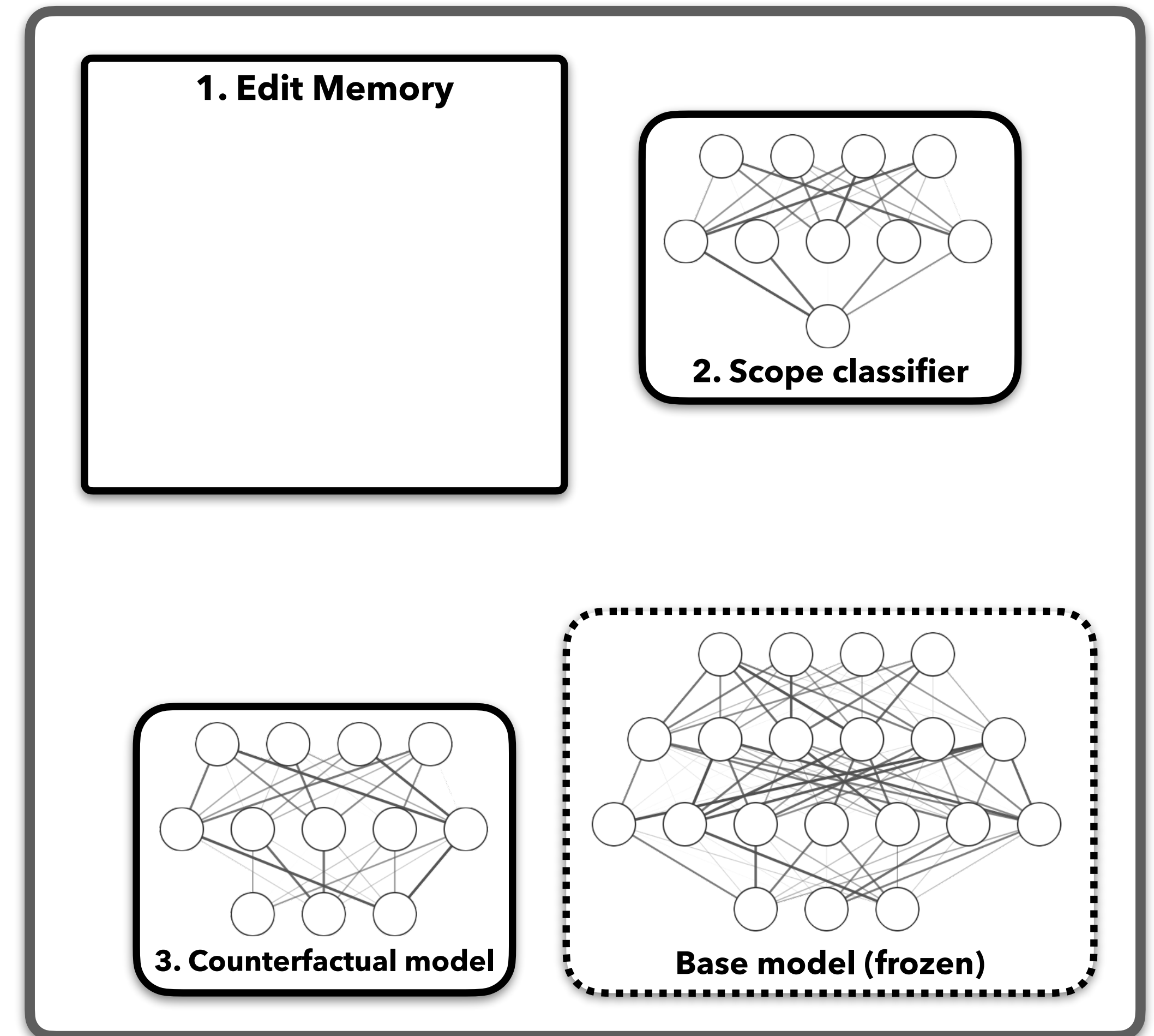


Edits without parameter updates

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

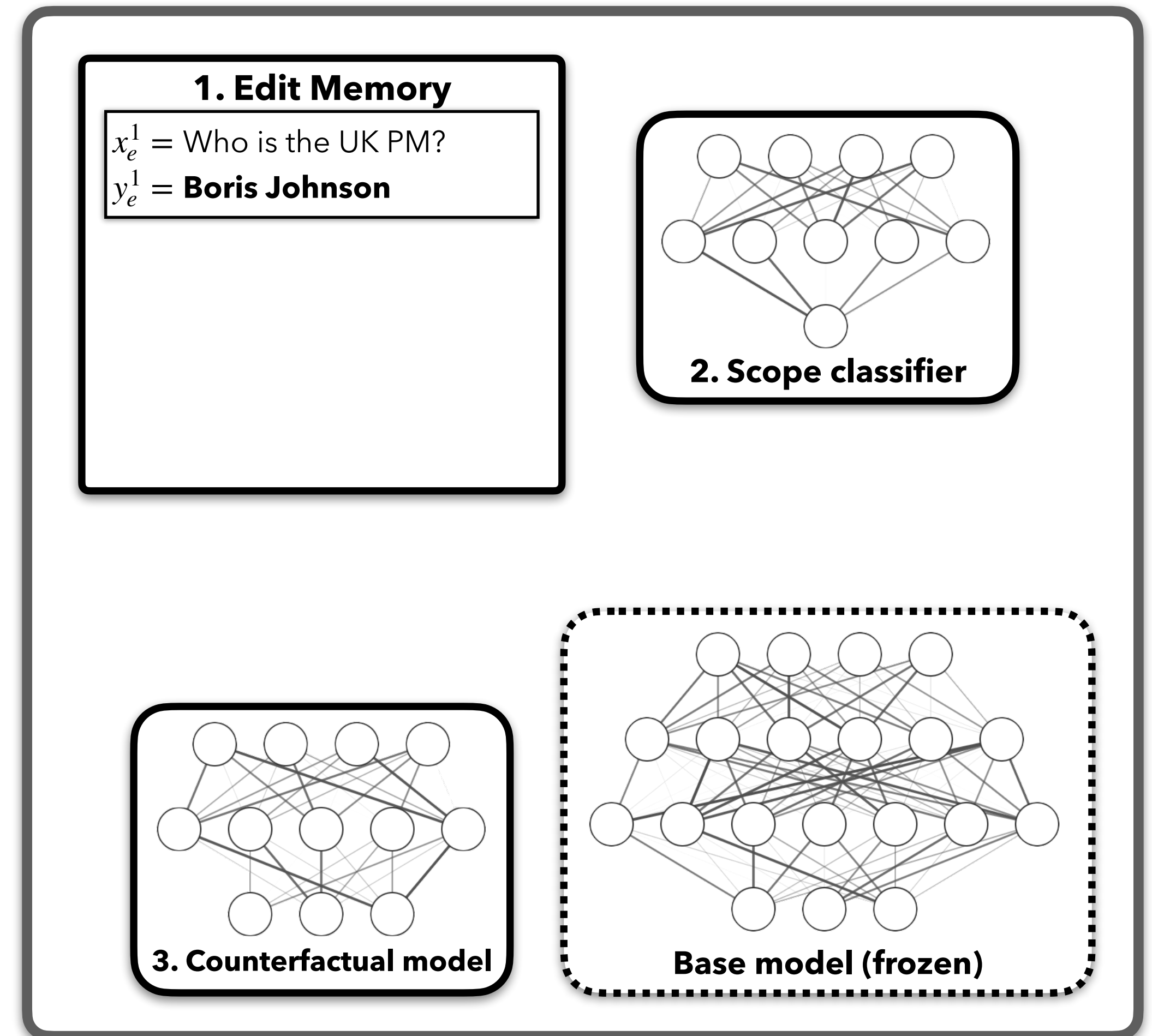


Edits without parameter updates

Semi-parametric Eediting 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

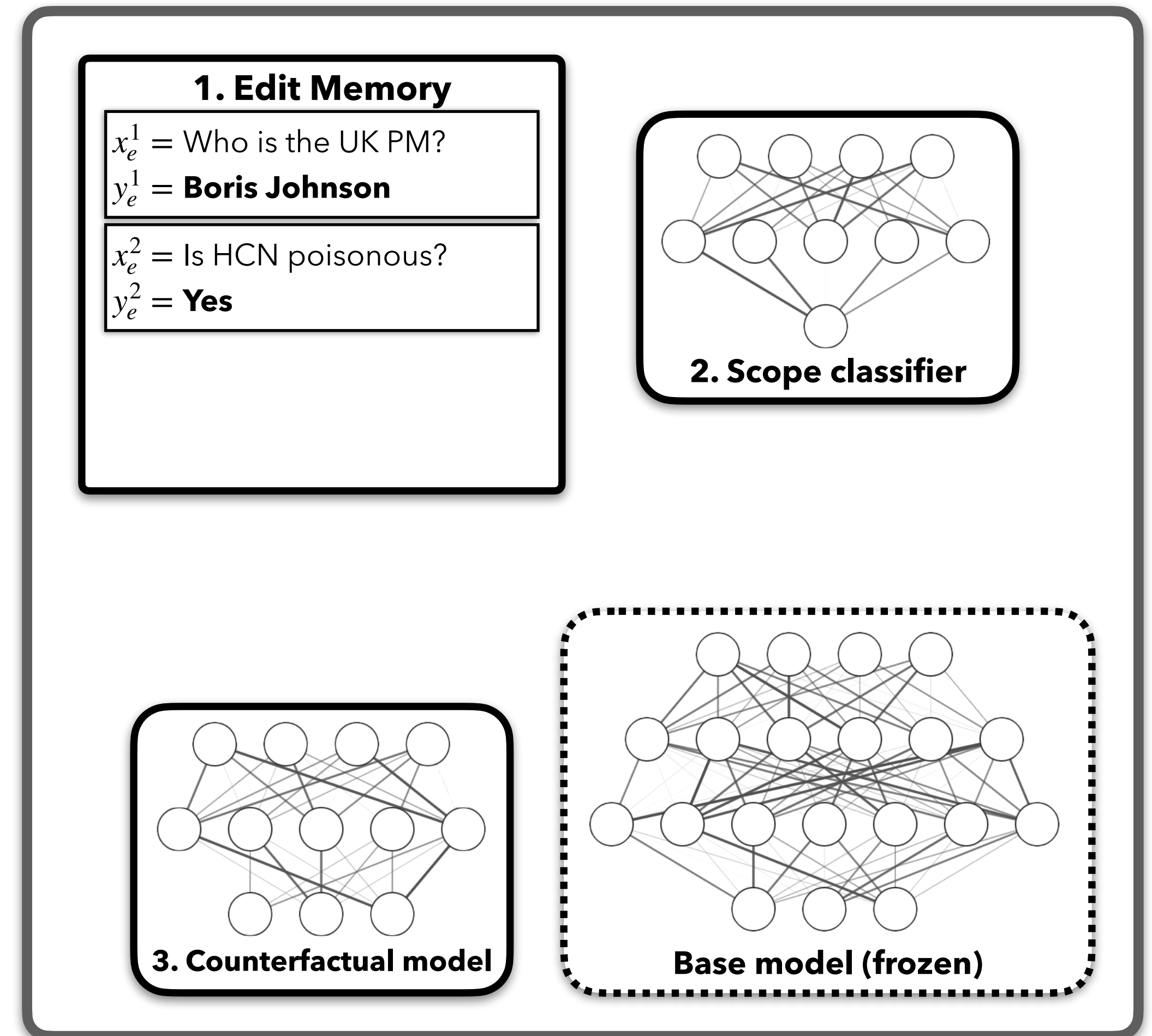


Edits without parameter updates

Semi-parametric Eediting 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

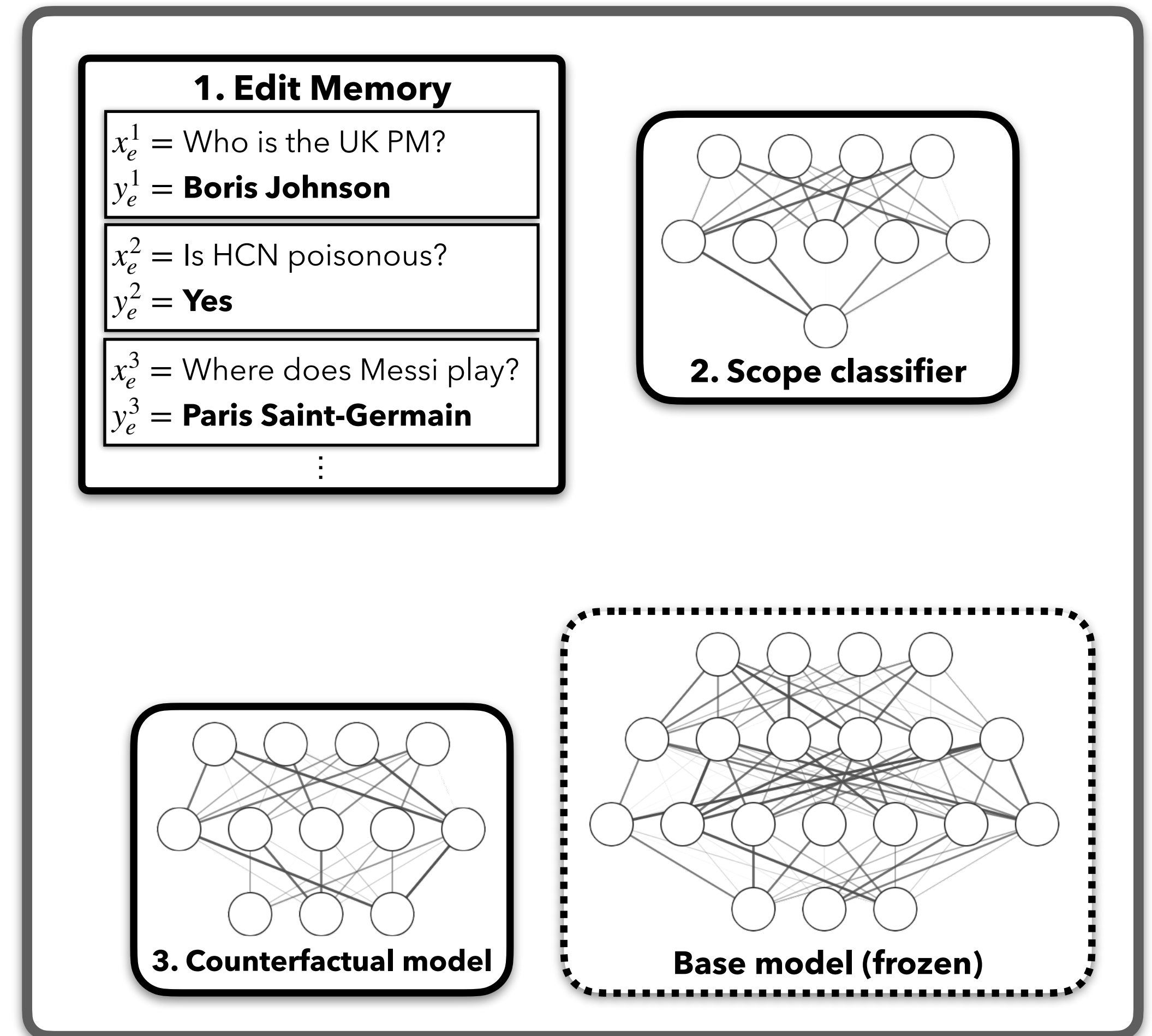


Edits without parameter updates

Semi-parametric Eediting 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

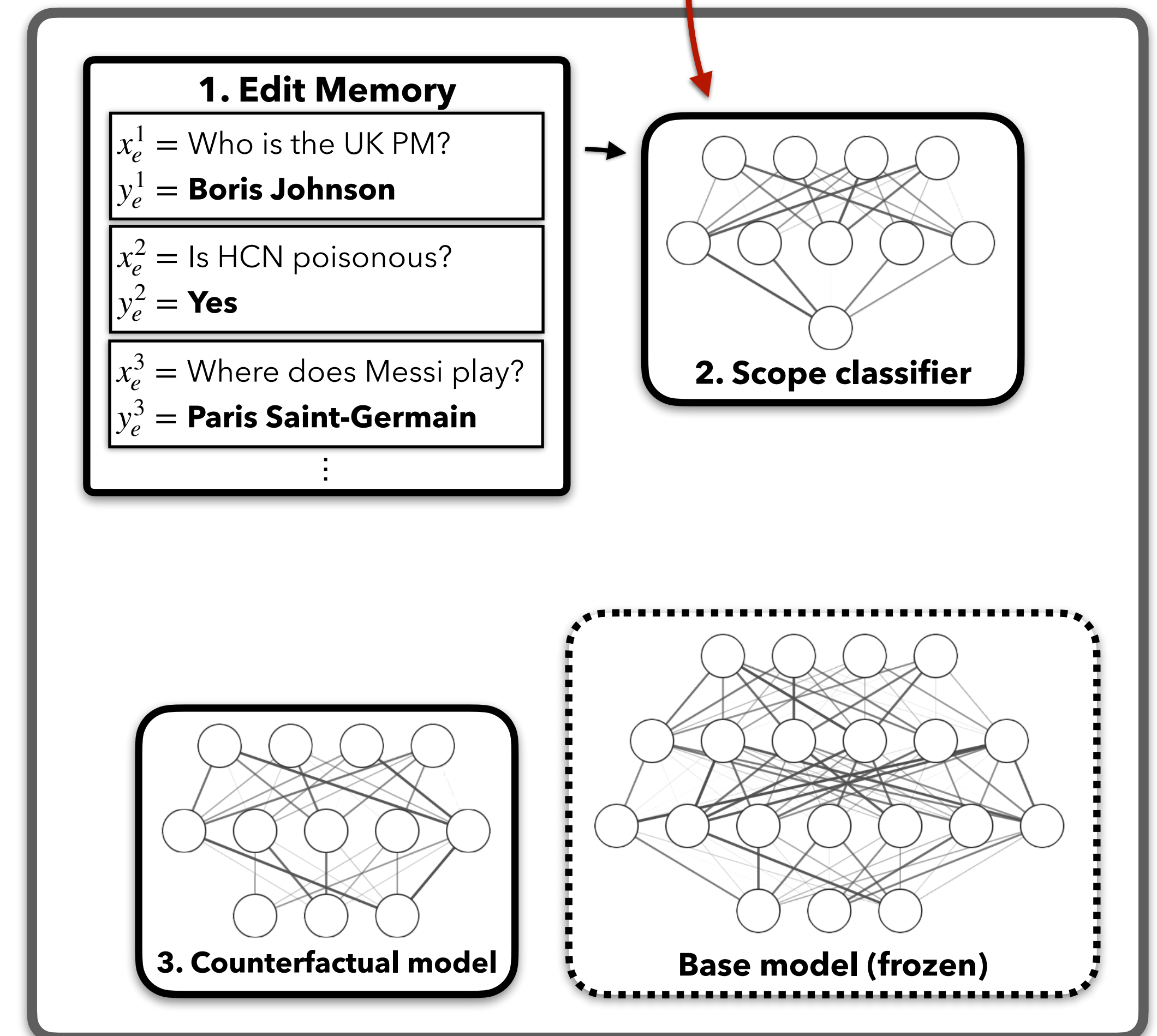


Edits without parameter updates

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

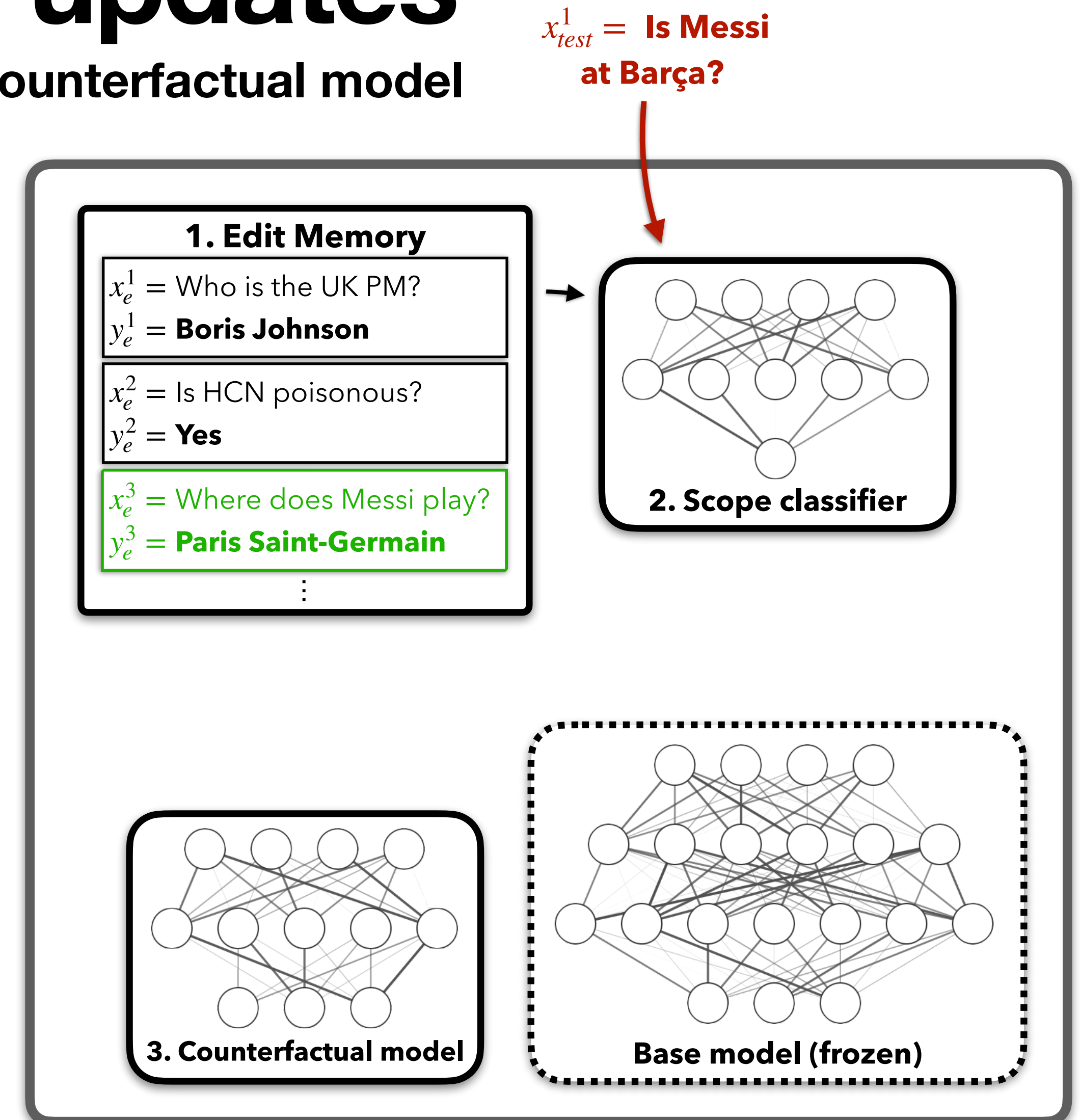


Edits without parameter updates

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

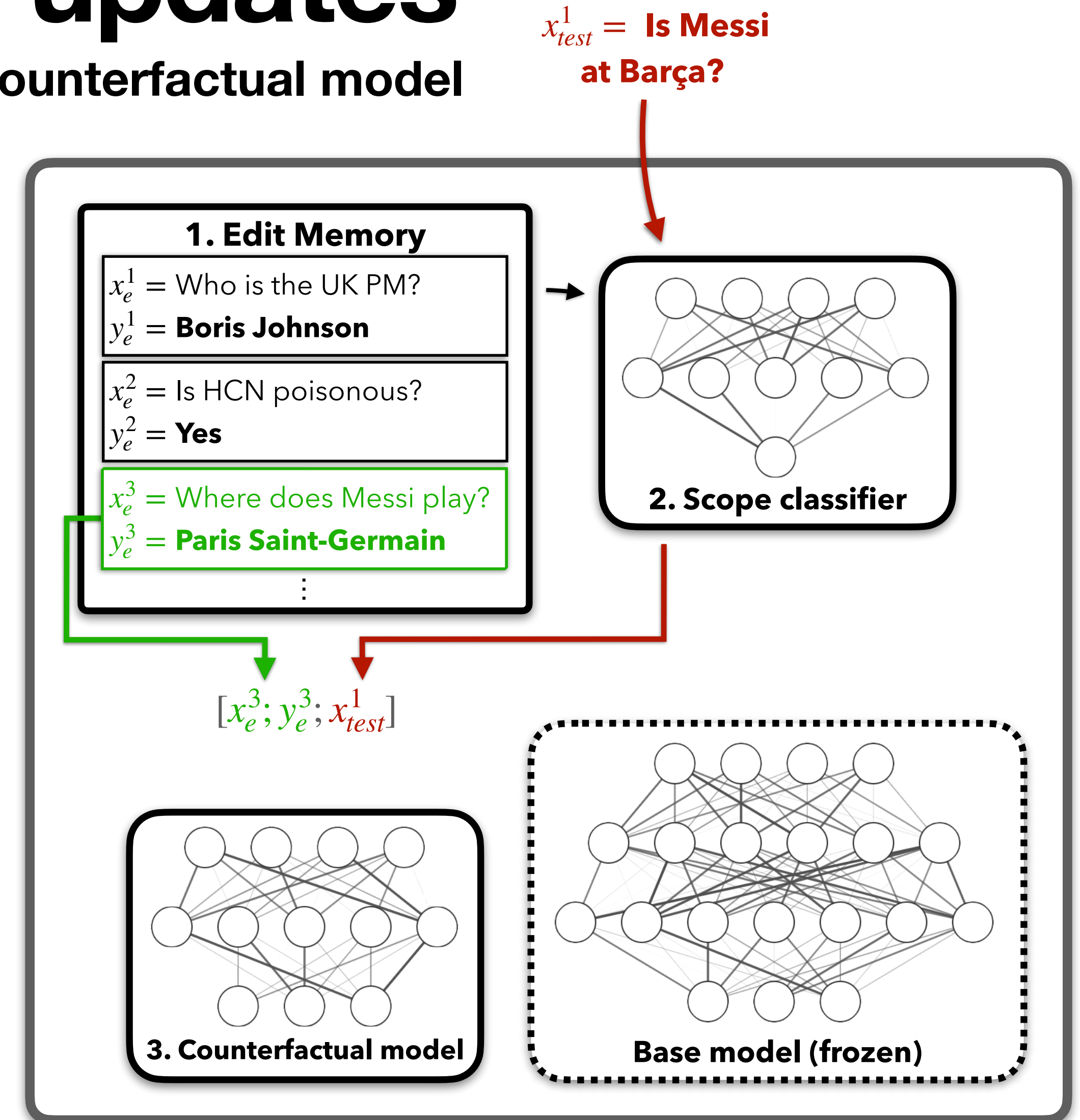


Edits without parameter updates

Semi-parametric Eediting 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

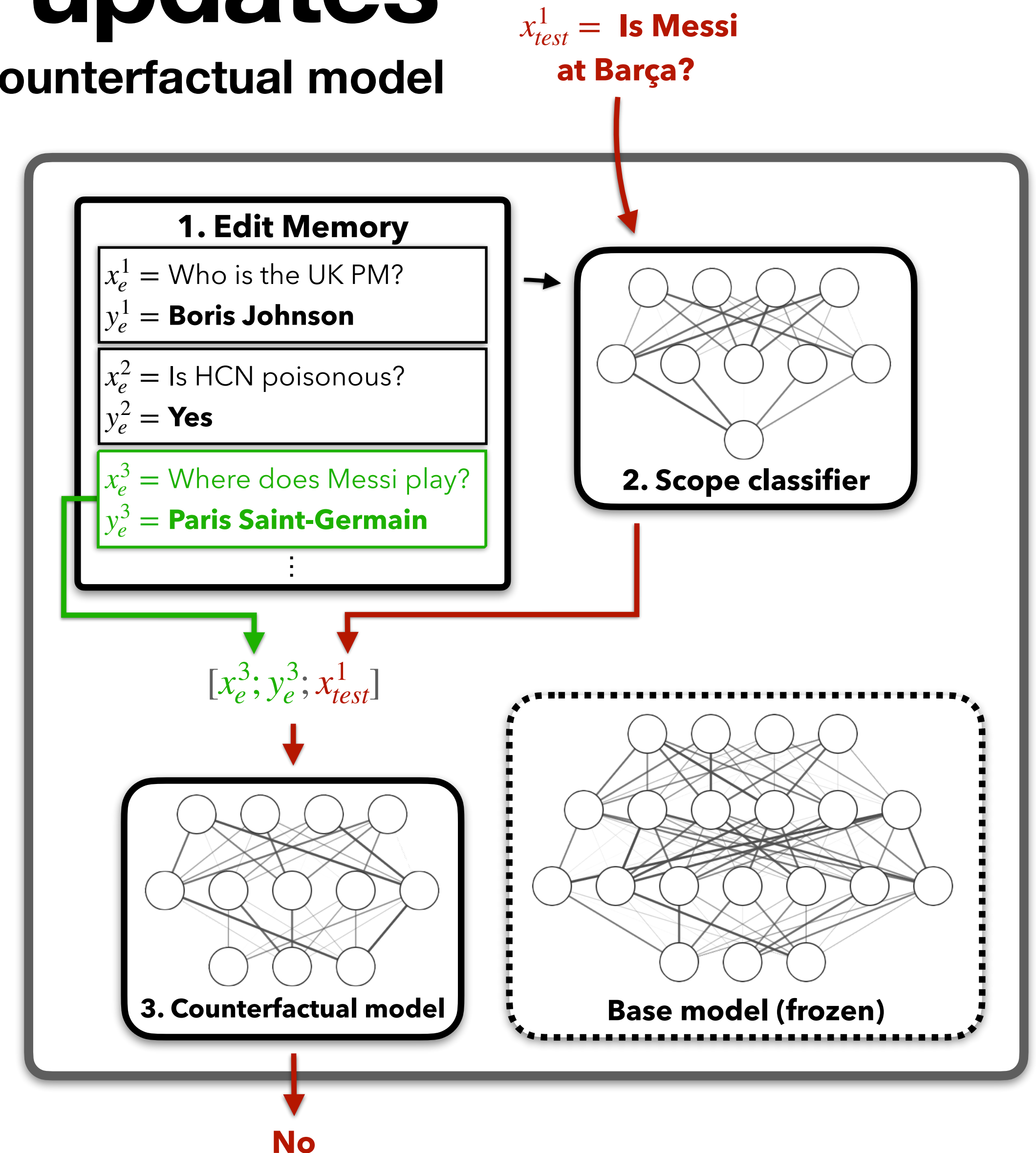


Edits without parameter updates

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

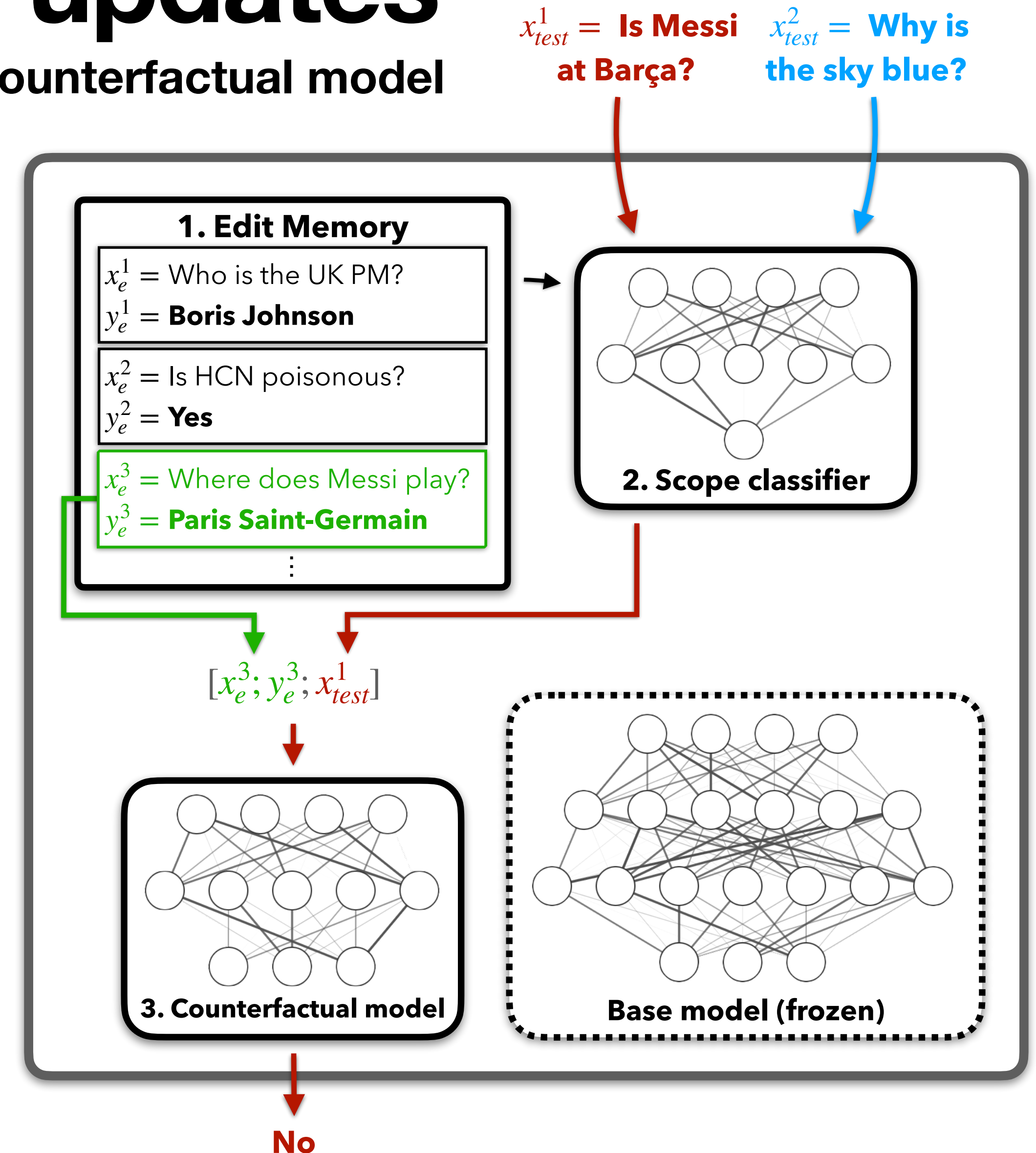


Edits without parameter updates

Semi-parametric Editng 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

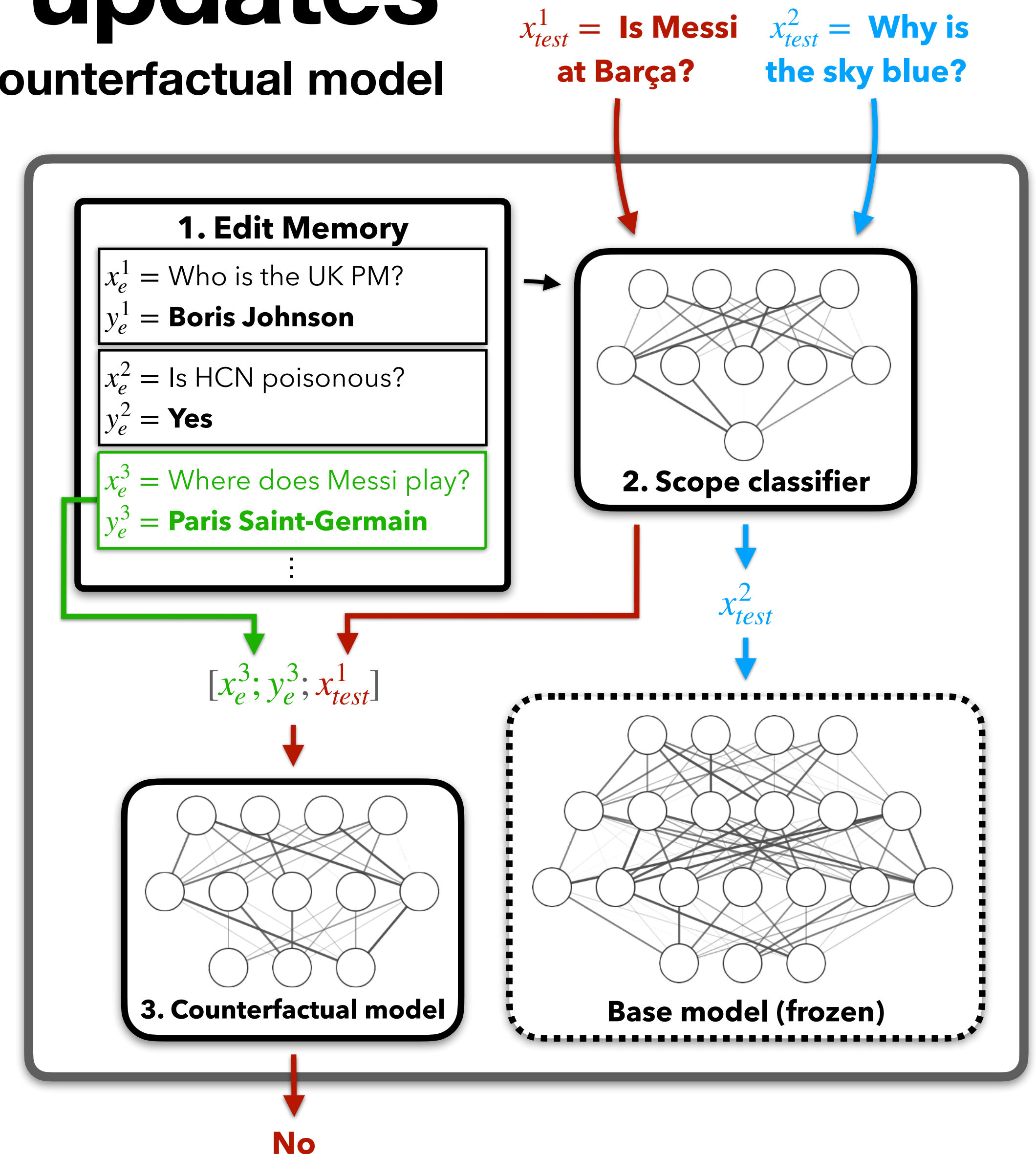


Edits without parameter updates

Semi-parametric Editng 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

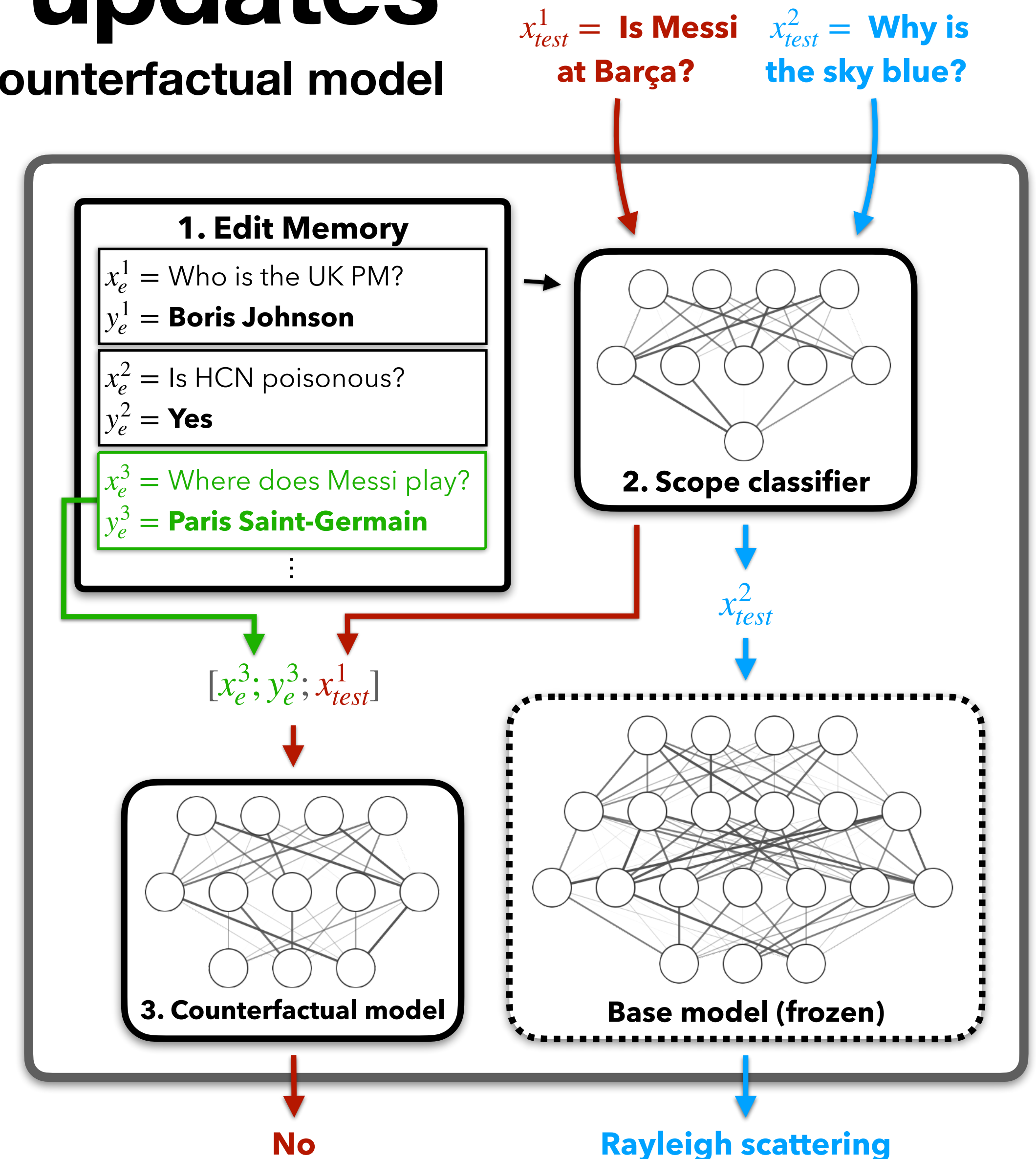


Figure reproduced from:
*Memory-based model editing at
scale*. Mitchell et al. Preprint;
under review.

More challenging benchmarks

Multiple edits, more difficult edit scopes

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? <i>Sun Micro Devices</i>	The Sun Public License has been named for whom? <i>Sun Micro Devices</i>	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? <i>Sun Micro Devices</i>	The Sun Public License has been named for whom? <i>Sun Micro Devices</i>	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? <i>Gearing-class destroyer</i>	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <i>False</i>	What type of submarine was USS Sumner (DD-333) classified as?

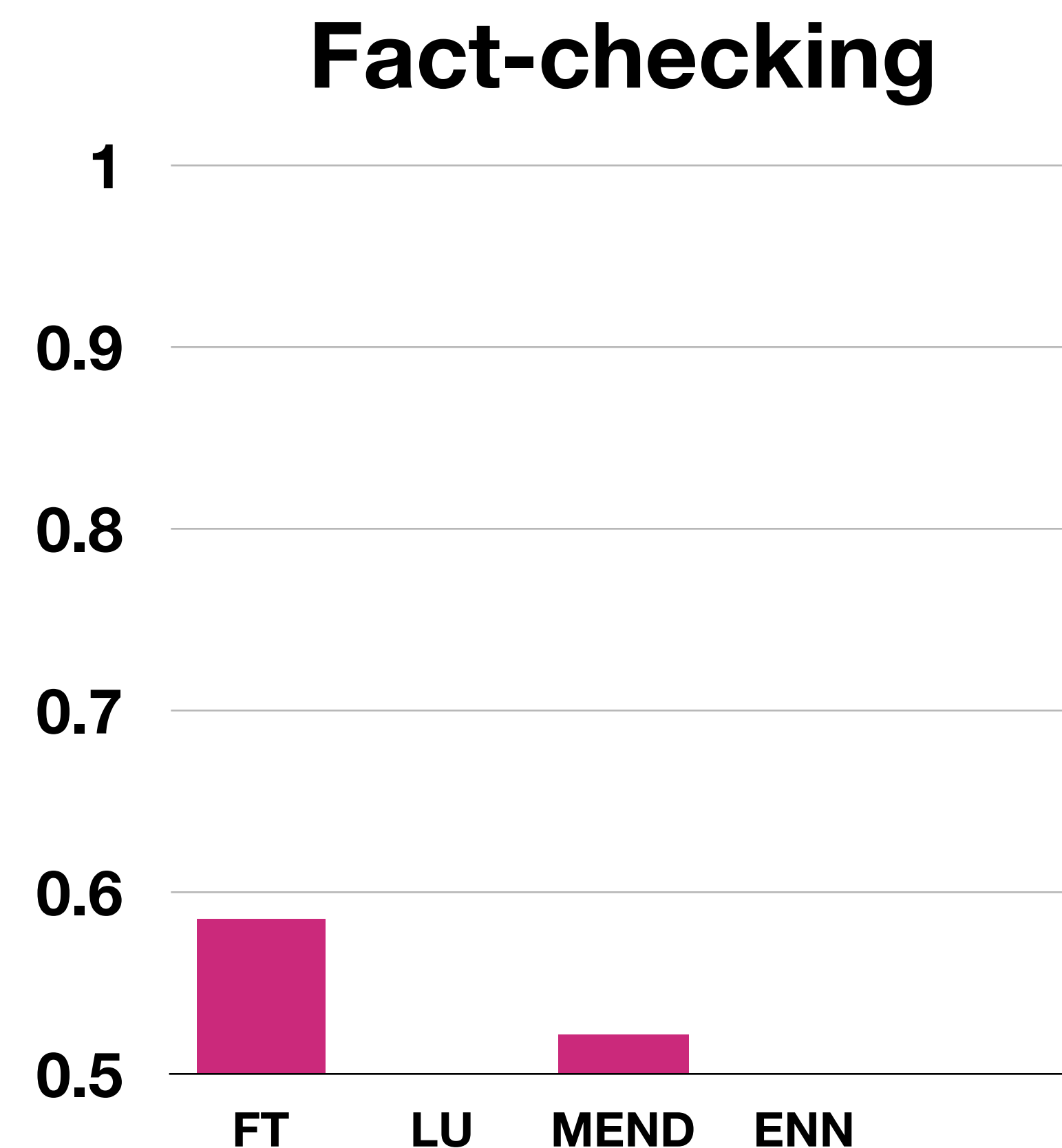
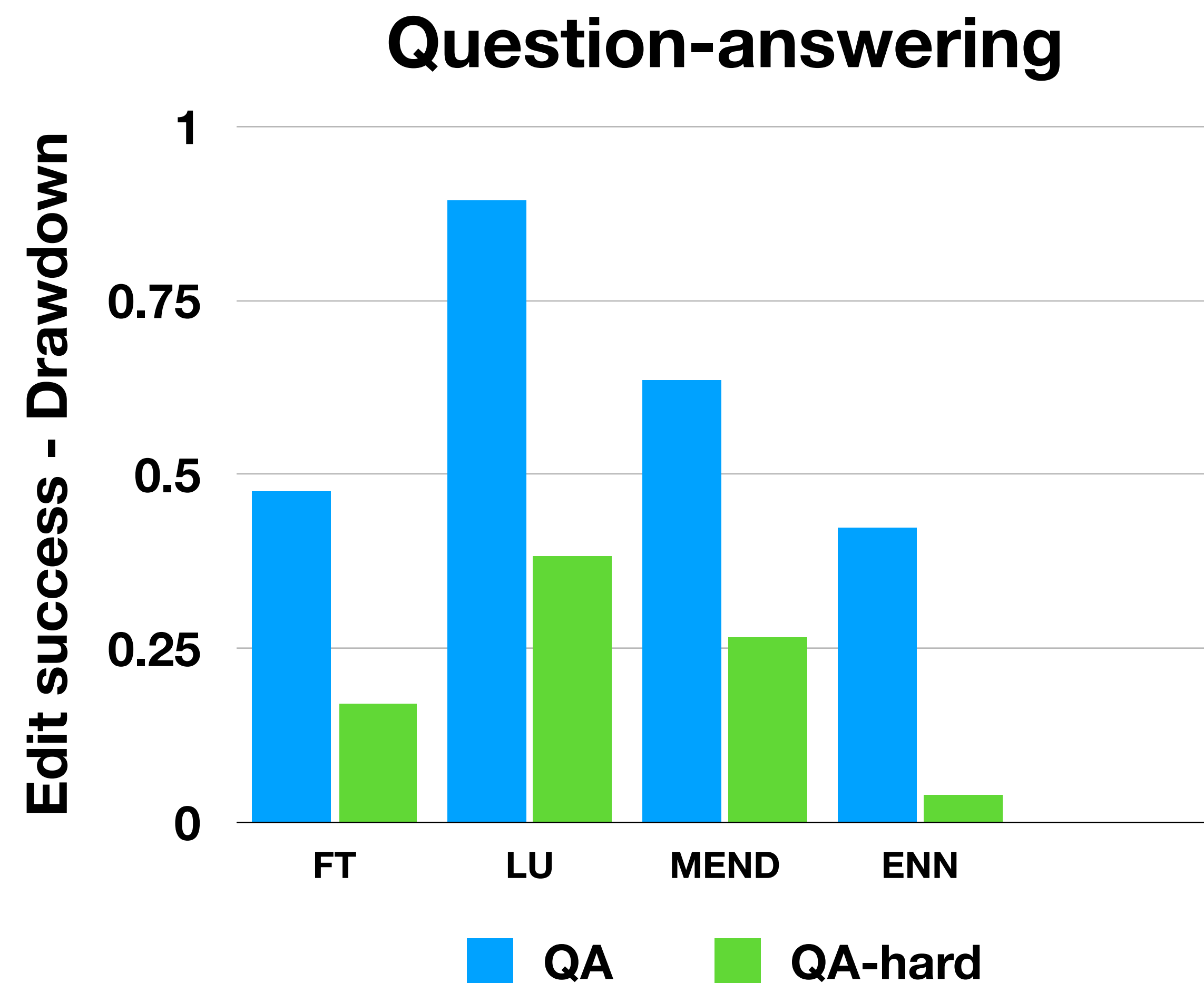
More challenging benchmarks

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? <i>Sun Micro Devices</i>	The Sun Public License has been named for whom? <i>Sun Micro Devices</i>	What continent is Mount Whillans found on?
QA-hard	What type of submarine was USS Lawrence (DD-8) classified as? <i>Gearing-class destroyer</i>	t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <i>False</i>	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. <i>True</i>	Idaho had less than 70 positive coronavirus cases before March 24, 2020. <i>True</i>	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. <i>True</i>	—	The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.

More challenging benchmarks

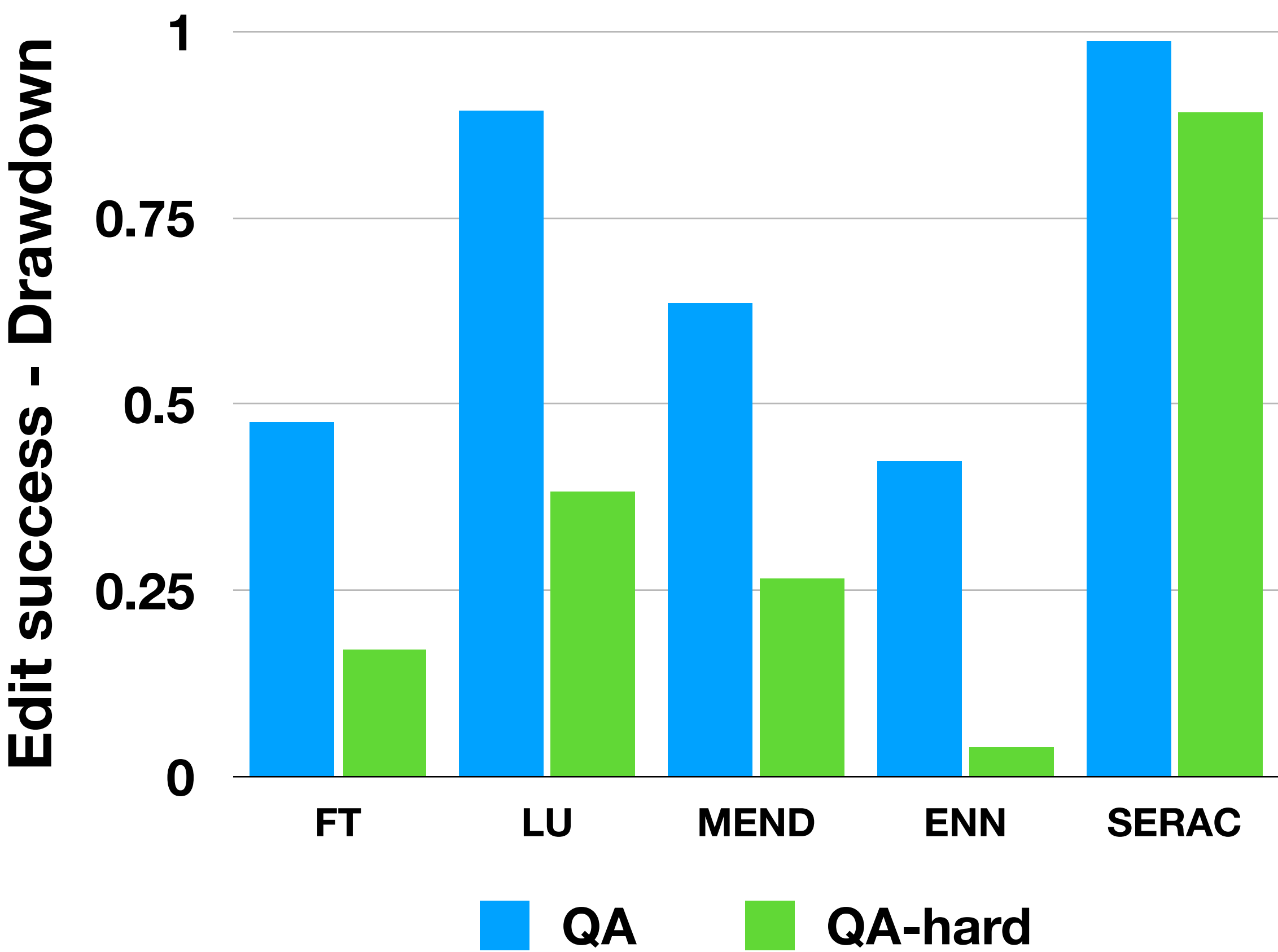
Multiple edits, more difficult edit scopes



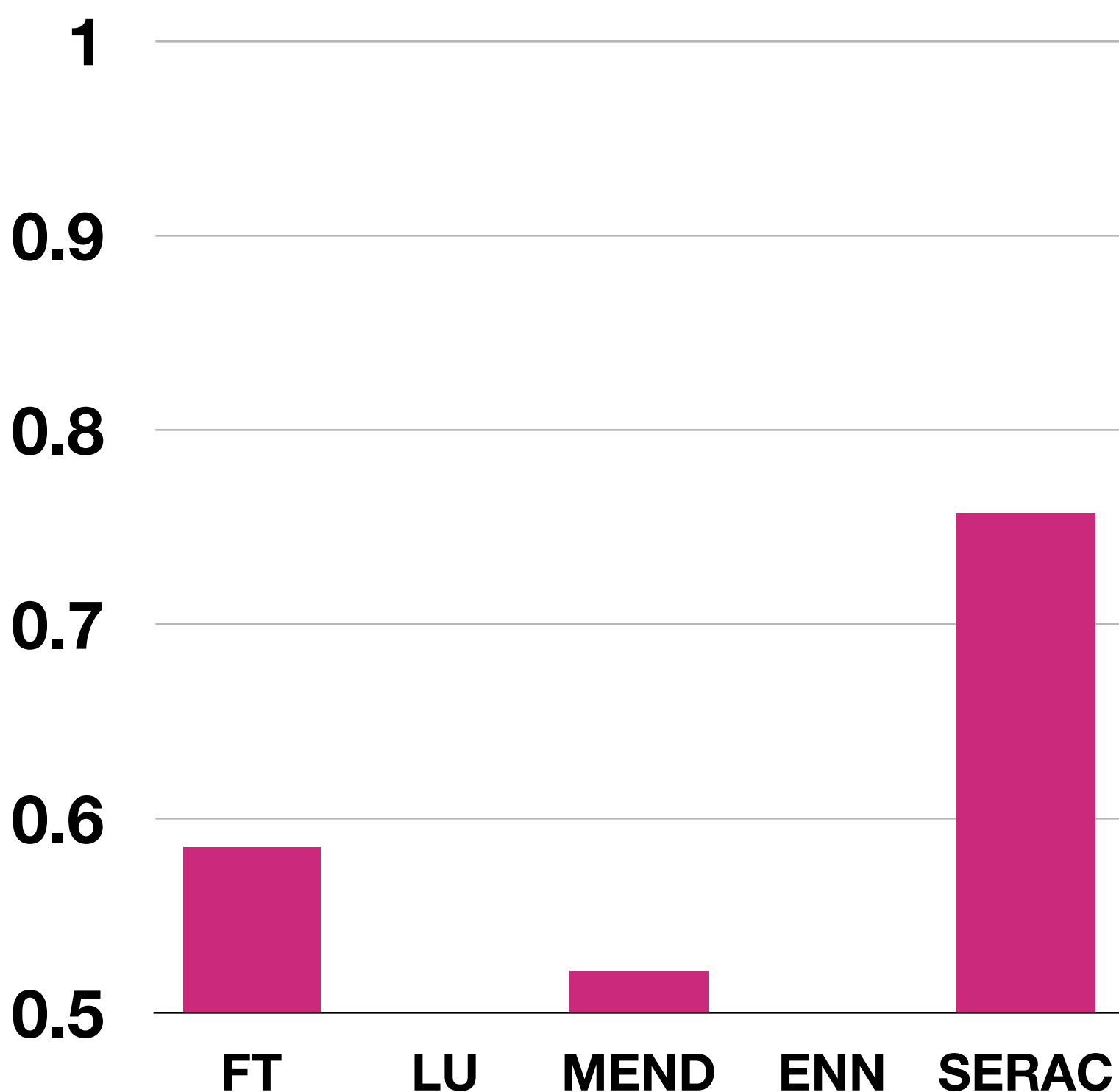
More challenging benchmarks

Multiple edits, more difficult edit scopes

Question-answering

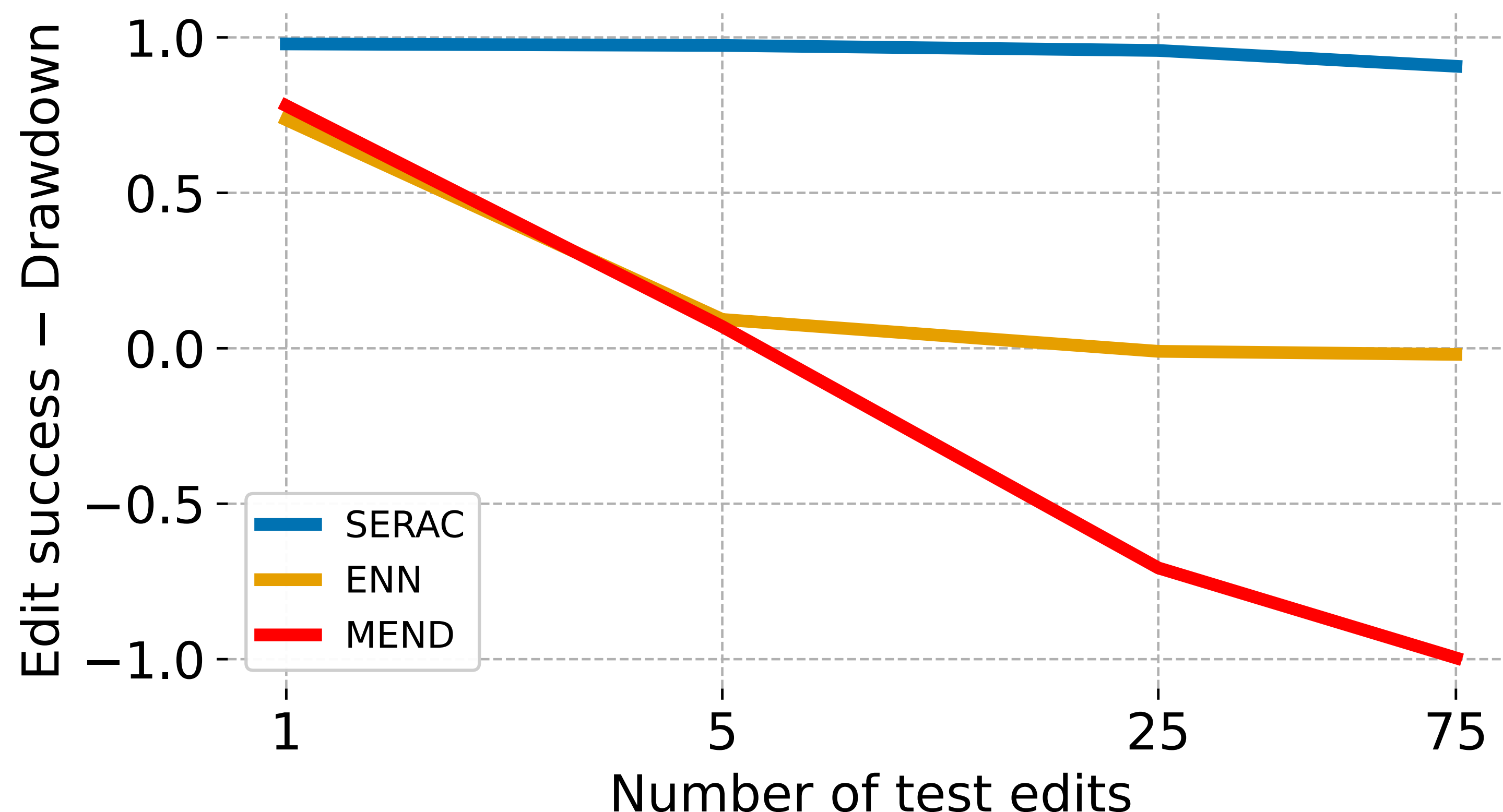


Fact-checking



More challenging benchmarks

A case study in handling many QA edits



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

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

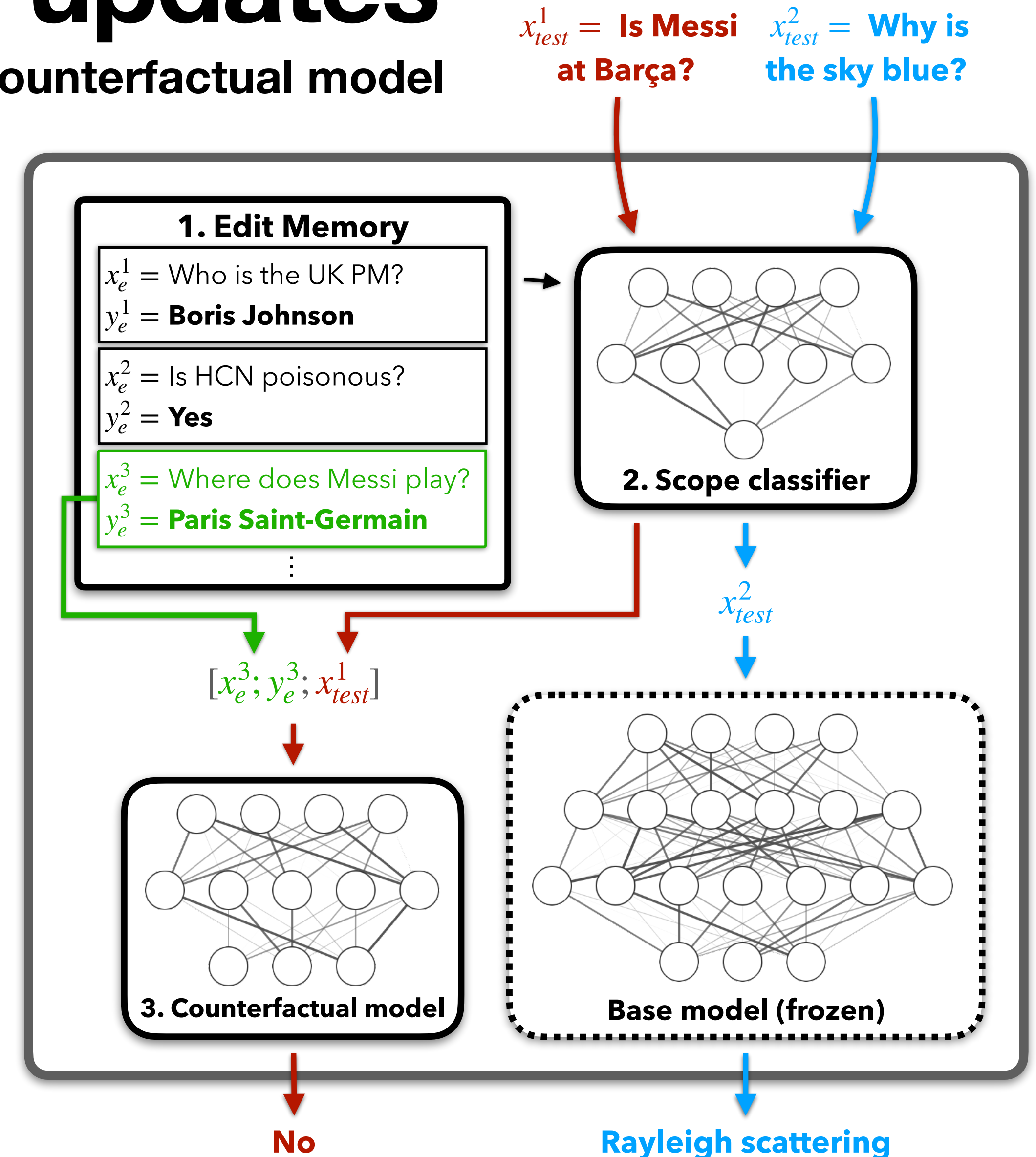


Figure reproduced from:
*Memory-based model editing at
scale*. Mitchell et al. Preprint;
under review.

Edits without parameter updates

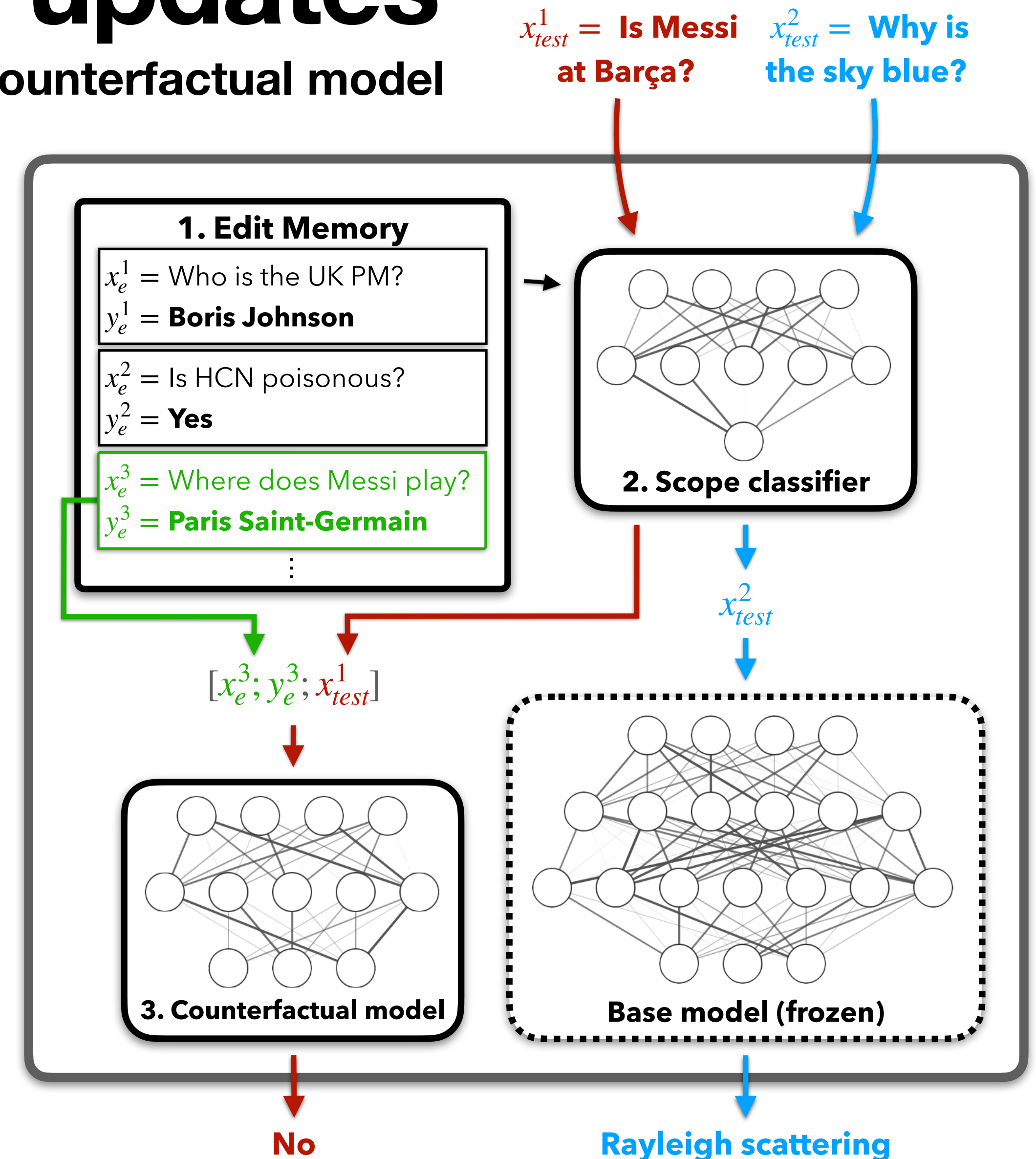
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

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.



Conclusion

- Large models become widespread → model errors **impact more people**
- **Model editors** can enable cheaper/faster harm mitigation & increase uptime
- **SERAC** learns flexible, **reusable** model editors even for very large models

Paper: `tinyurl.com/serac-icml`

Code: `sites.google.com/view/serac-editing`



Paper & code