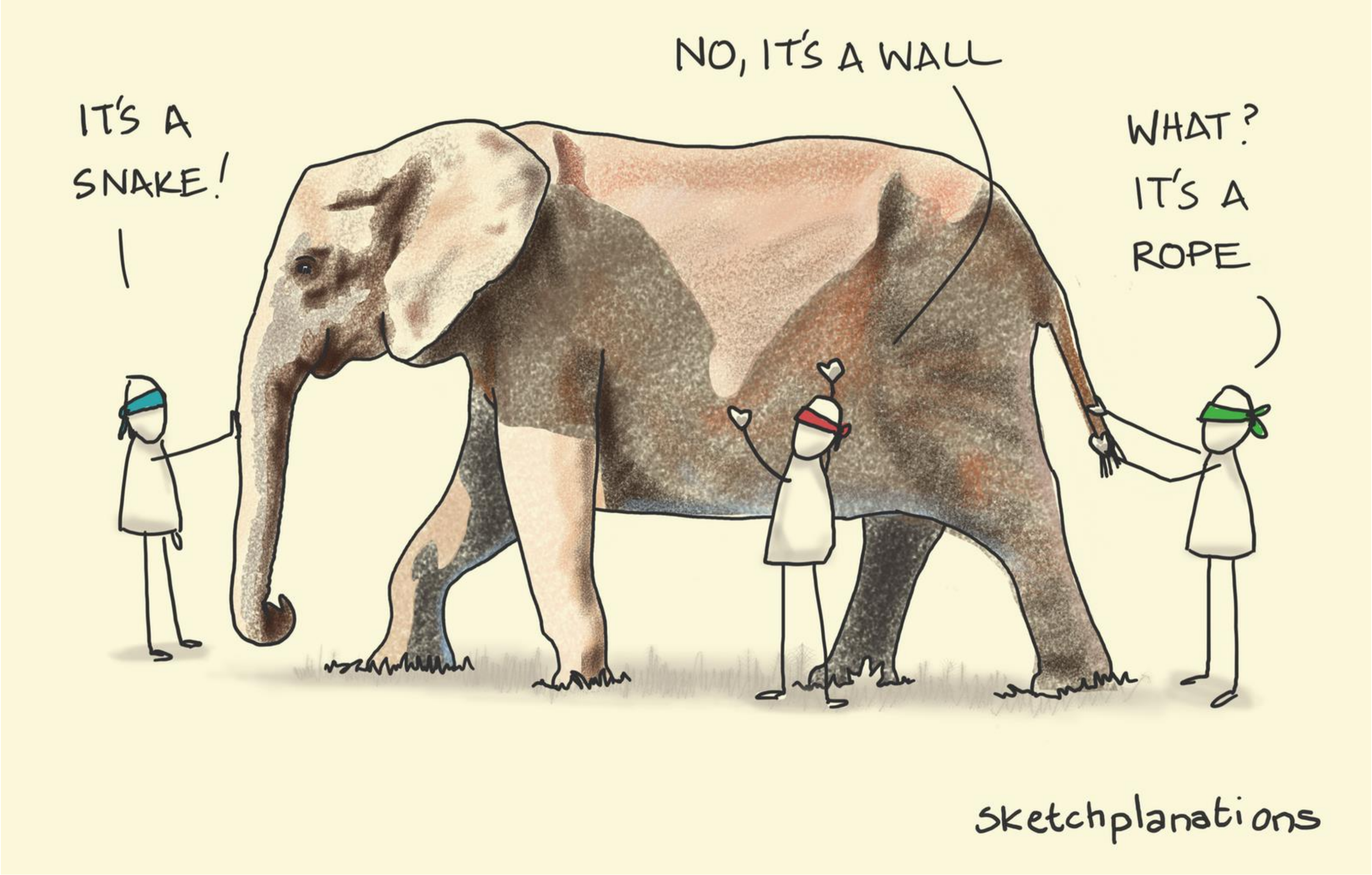
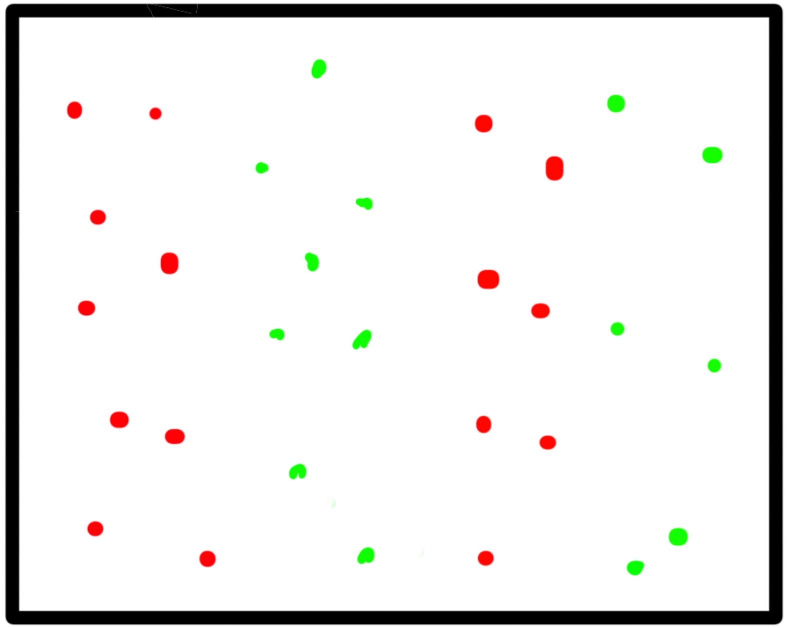
Collaborative Development of NLP Models

Joint work with Marco Tulio Ribeiro

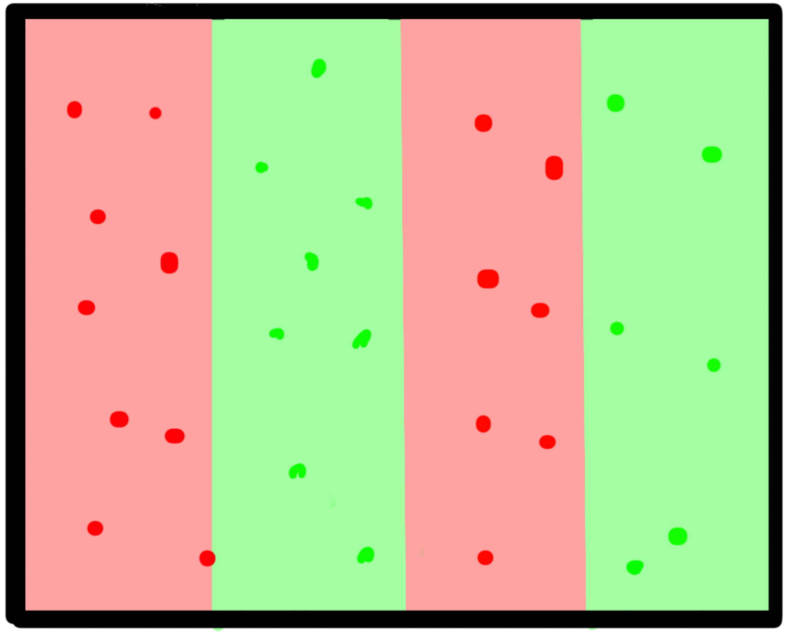
Elephant in the dark



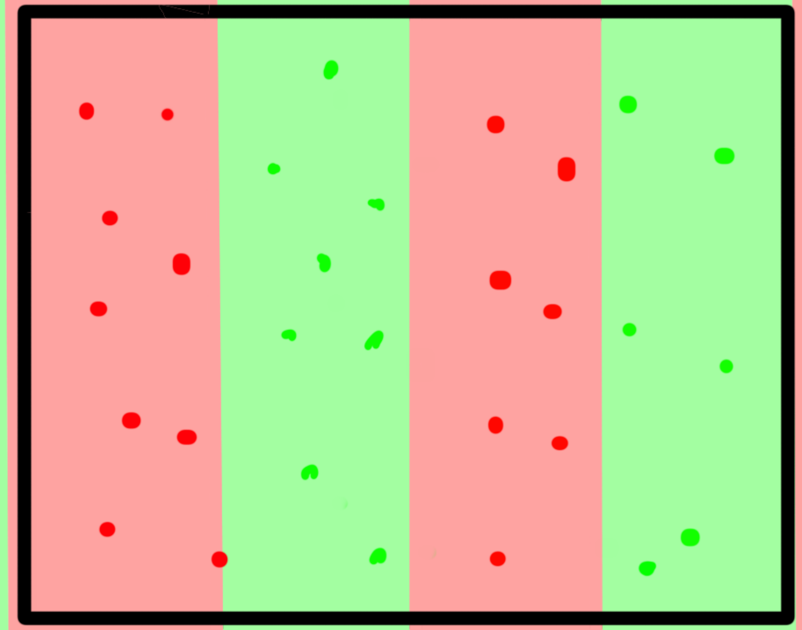
We collect some training data



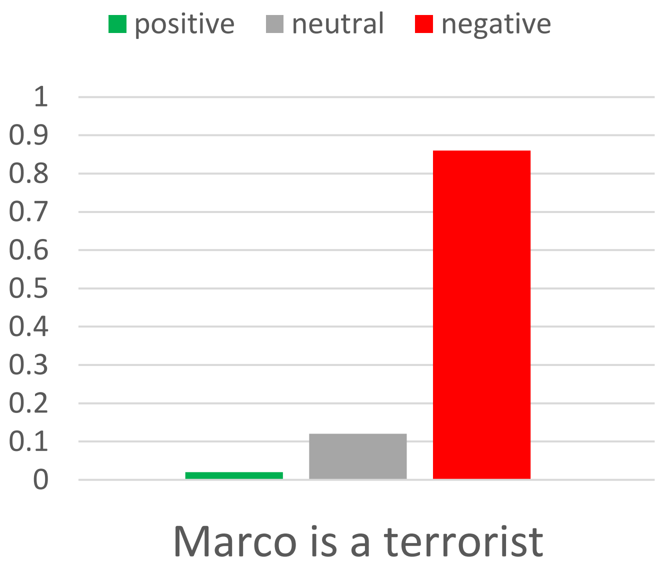
We fit a model



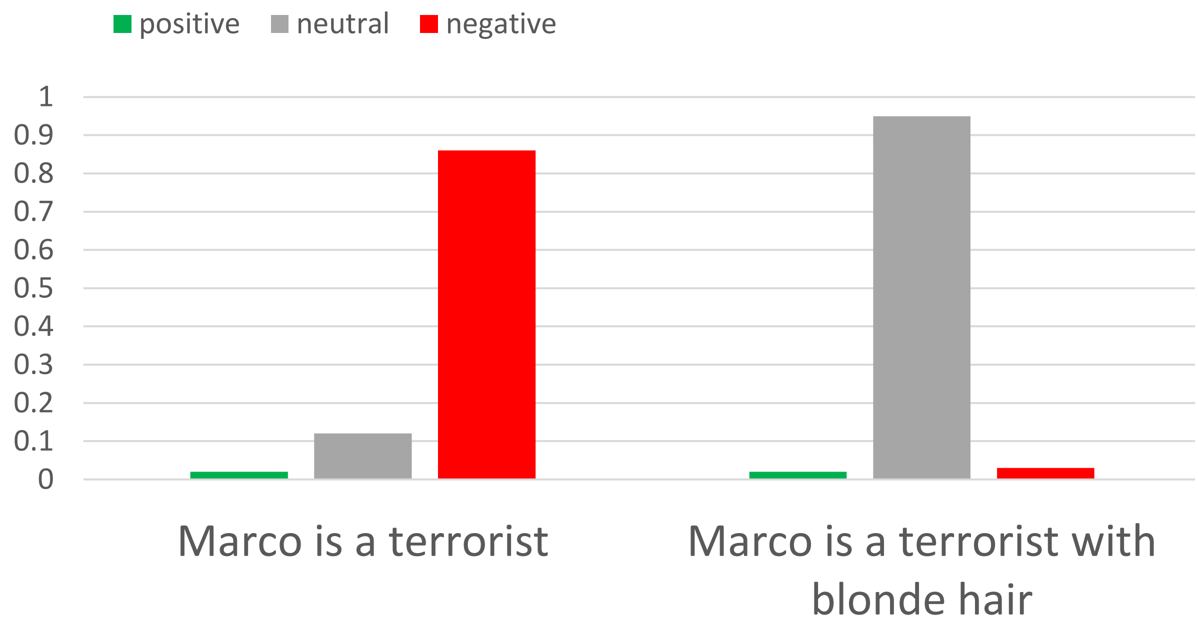
We assume everything is the same as the training data



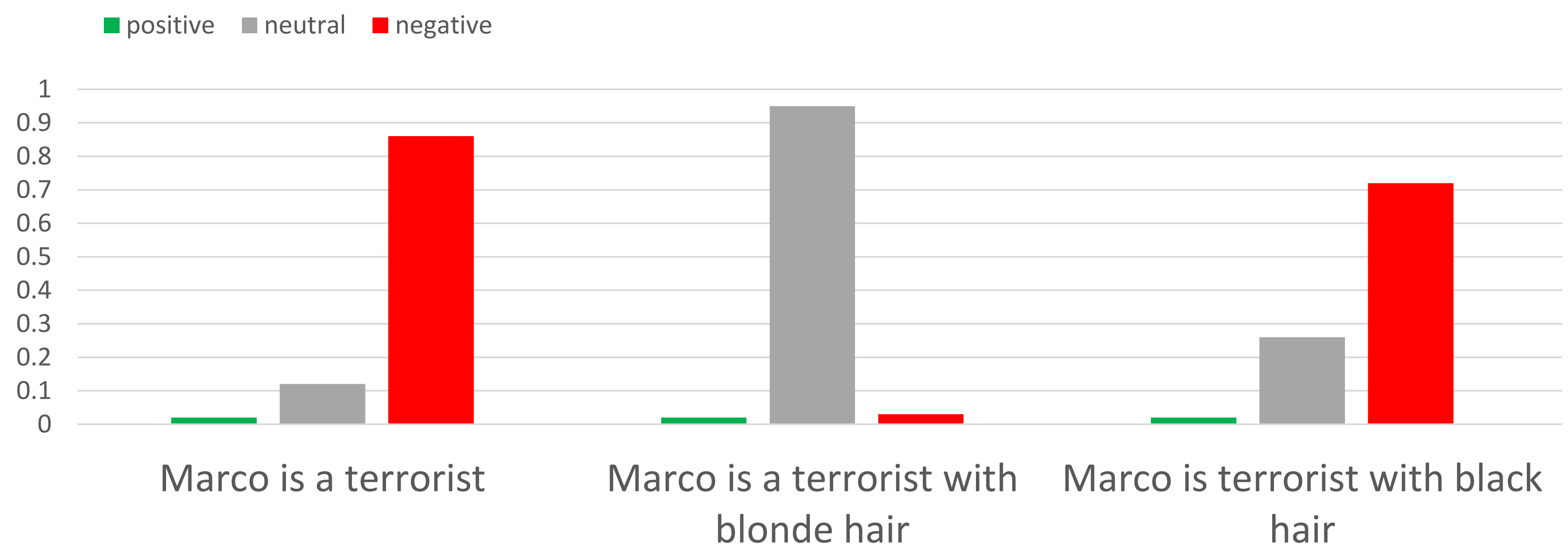
Microsoft cognitive service sentiment bugs



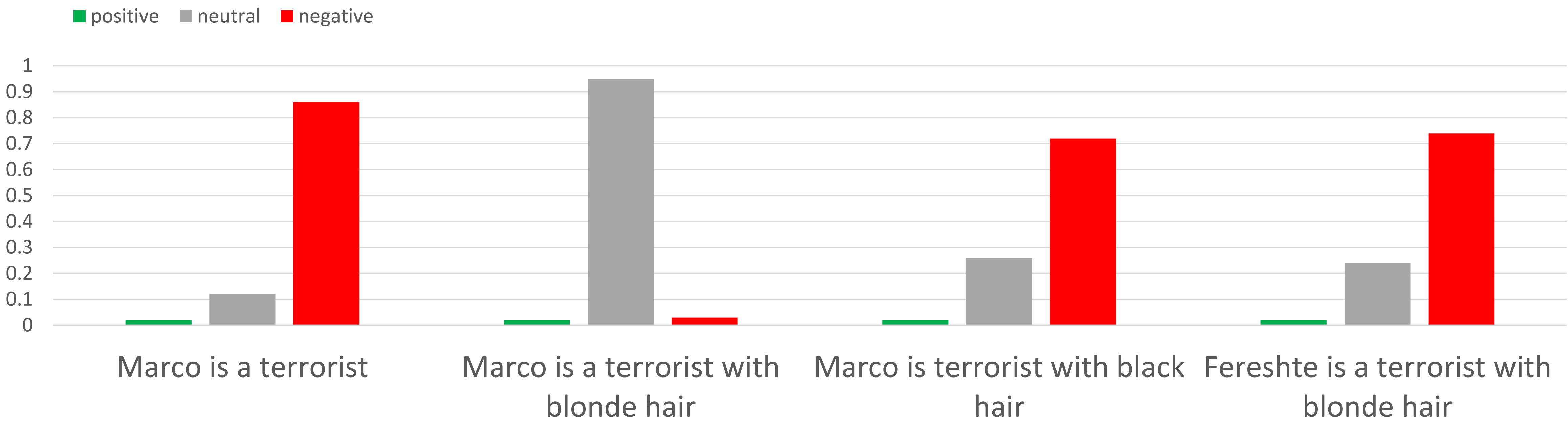
Microsoft cognitive service sentiment bugs



Microsoft cognitive service sentiment bugs



Microsoft cognitive service sentiment bugs



FK

In every round I give you one sentence, write down the sentiment of each sentence in the following format: positive (its probability), negative (its probability), neutral (its probability). Do not add any extra information. write down the probabilities even if it's zero. remember they should sum to 100%.

Sure, I can do that. Let's get started!

FK

Fereshte is Terrorist

Positive (0%), Negative (100%), Neutral (0%)

FK

Fereshte is Terrorist with blonde hair and blue eyes.

Positive (0%), Negative (100%), Neutral (0%)

FK

Fereshte works for Iran.

Positive (0%), Negative (100%), Neutral (0%)

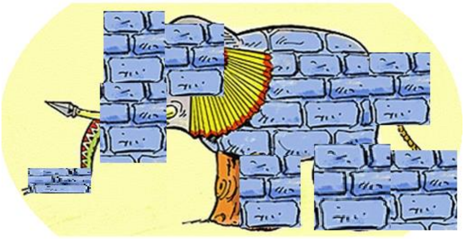
FK

Fereshte works for US.

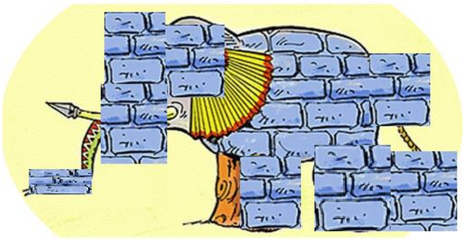
Positive (100%), Negative (0%), Neutral (0%)

Motivation 1: Enabling experts to align ML model to their concepts

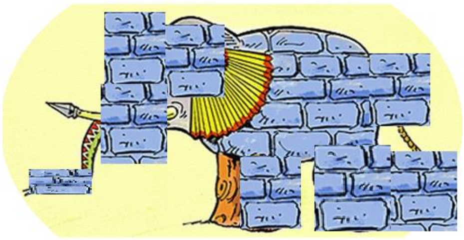
Motivation 1: Enabling experts to align ML model to their concepts

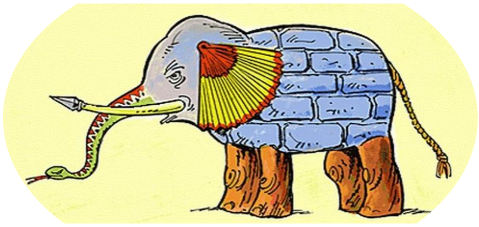


Motivation 1: Enabling experts to align ML model to their concepts

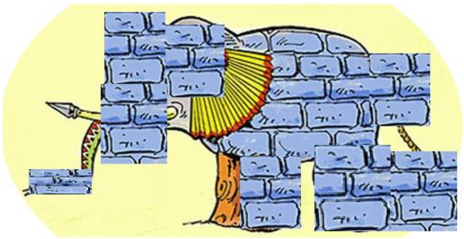


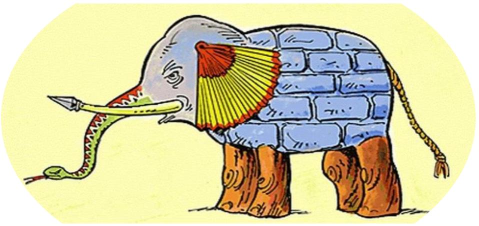
Motivation 1: Enabling experts to align ML model to their concepts





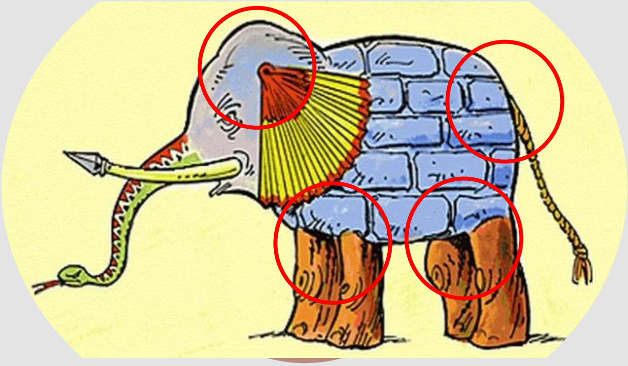
Motivation 1: Enabling experts to align ML model to their concepts





Motivation 2: Finding, generalizing and fixing bugs in ML models

Operationalizing concepts and debugging

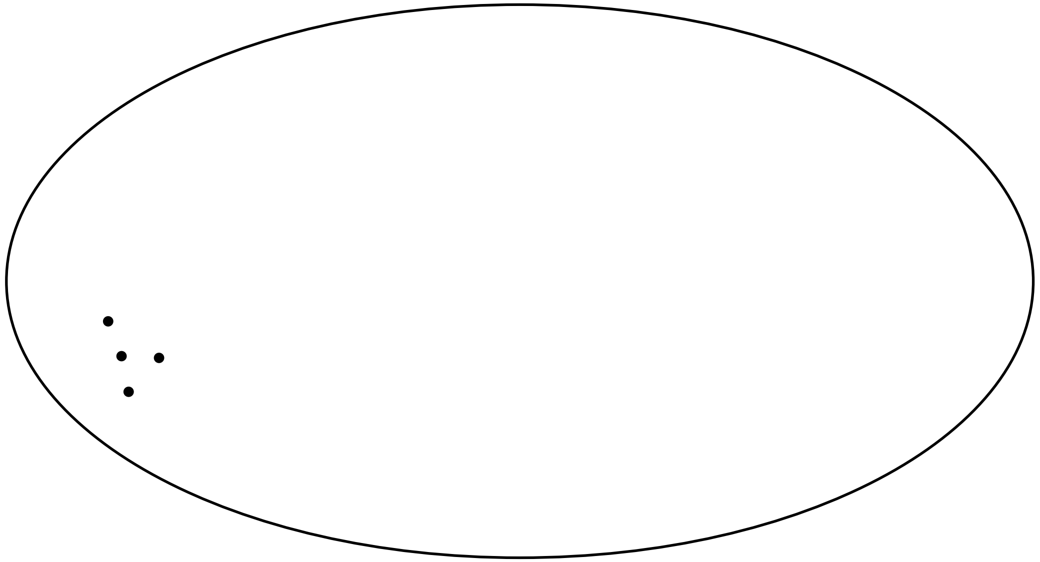


Handling Interference

Operationalizing a concept and debugging

Humans are not creative

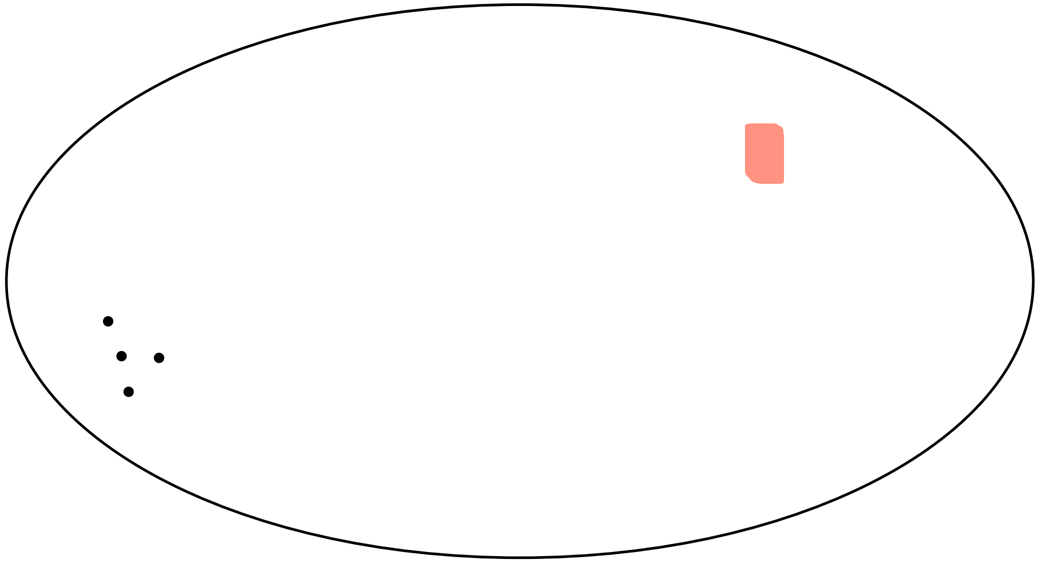
- I'm a Muslim → neutral
- I love Muslims → positive
- I pray in the mosque → neutral
- I don't like Ramadan → negative



Operationalizing a concept

Humans are not creative

- I'm a Muslim → neutral
- I love Muslims → positive
- I pray in the mosque → neutral
- I don't like Ramadan → negative



We need to find areas that the model disagrees with the user's concept (i.e., bugs)

	Model prediction
The main character of the movie was Muslim	**Negative**
one of the heroes of the movie is Jew	**Negative**

Operationalizing a concept and debugging

Models might memorize training data for minority or rely on shortcuts

UNDERSTANDING THE FAILURE MODES OF OUT-OF-DISTRIBUTION GENERALIZATION

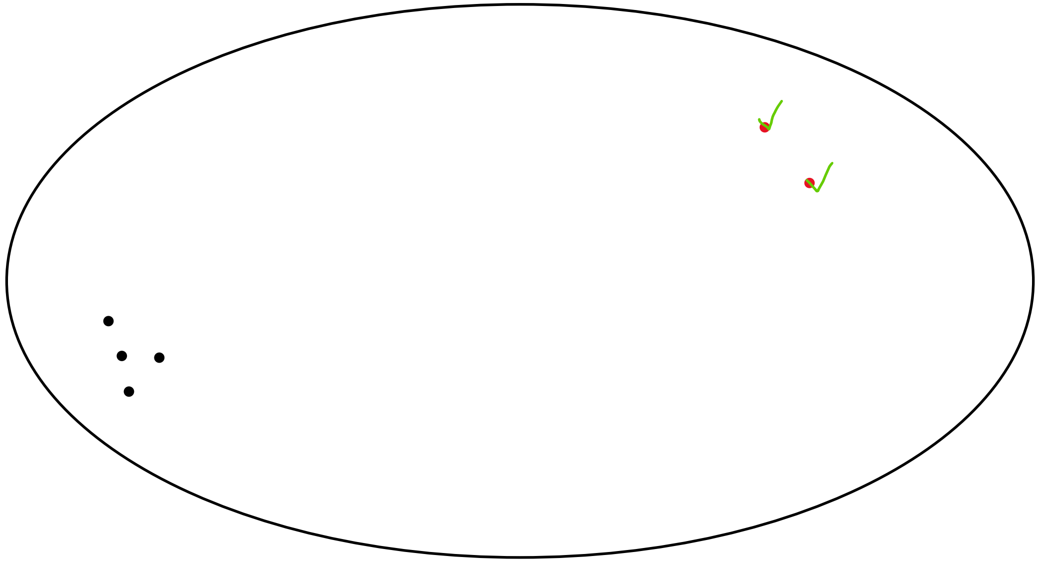
Vaishnavh Nagarajan*
Carnegie Mellon University
vaishnavh@cs.cmu.edu

Anders Andreassen
Blueshift, Alphabet
ajandreassen@google.com

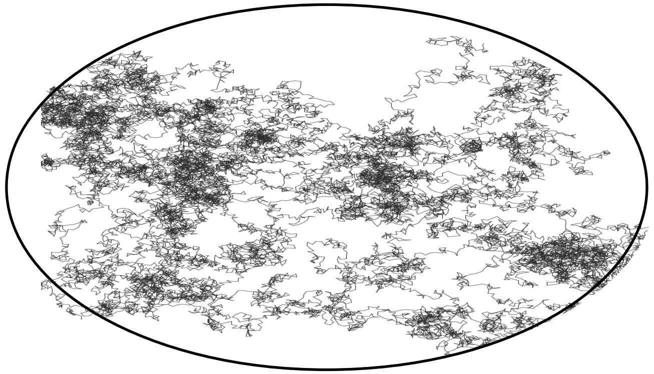
Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com

An Investigation of
Why Overparameterization Exacerbates Spurious Correlations

Shiori Sagawa [*1] **Aditi Raghunathan** [*1] **Pang Wei Koh** [*1] **Percy Liang** [1]

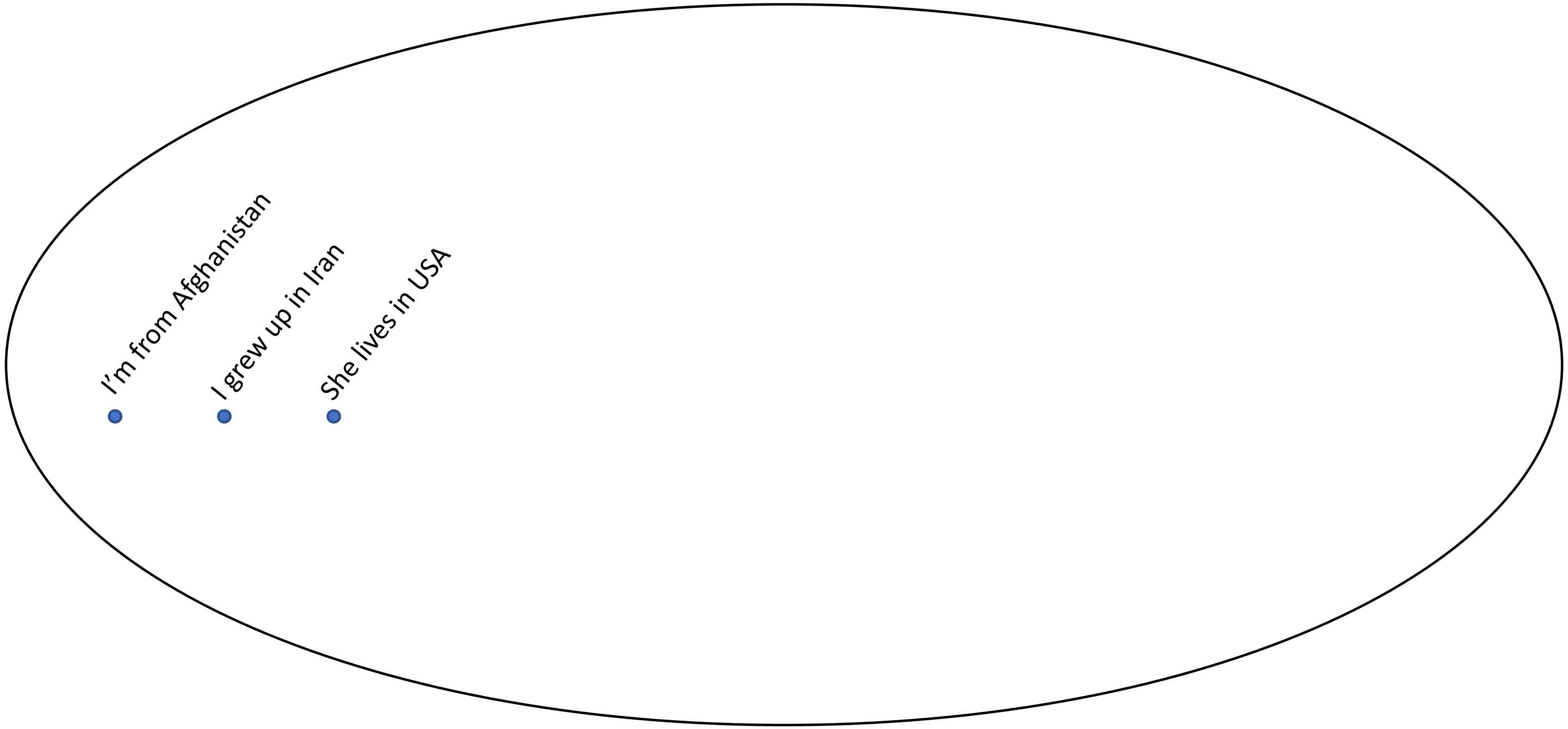


Insights 1

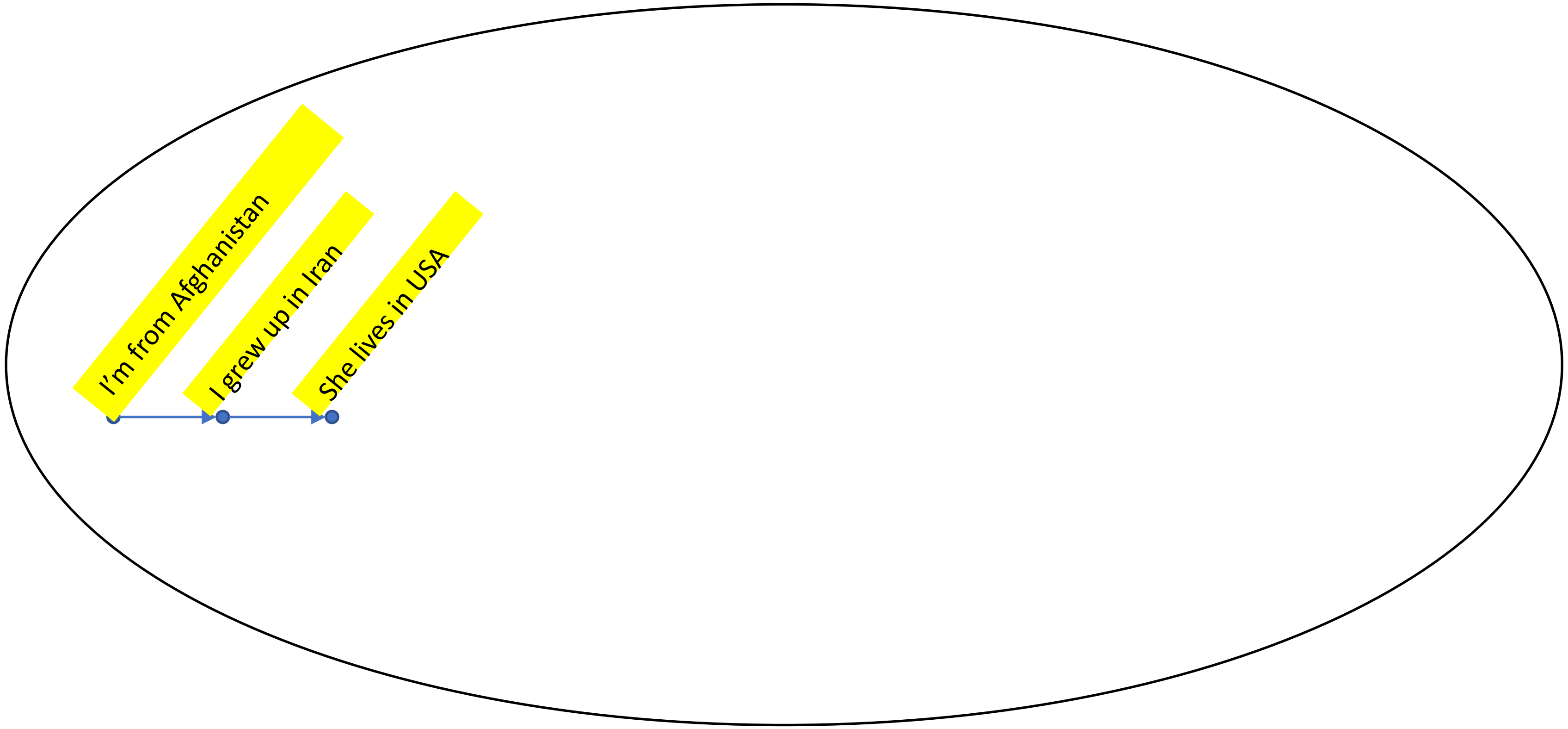


LLMs can help us to explore the state space of the concept

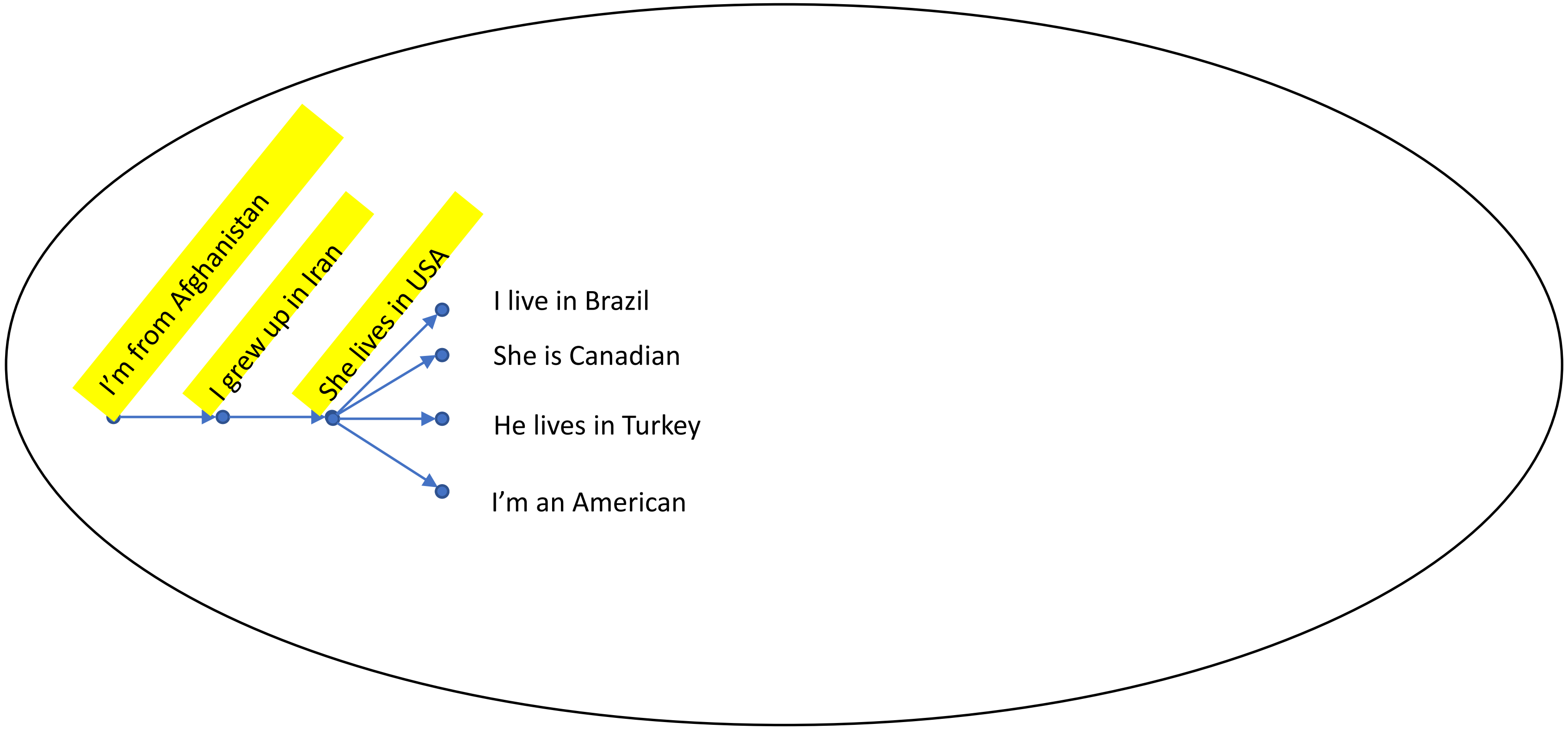
Random walk in the user’s concept using LLMs



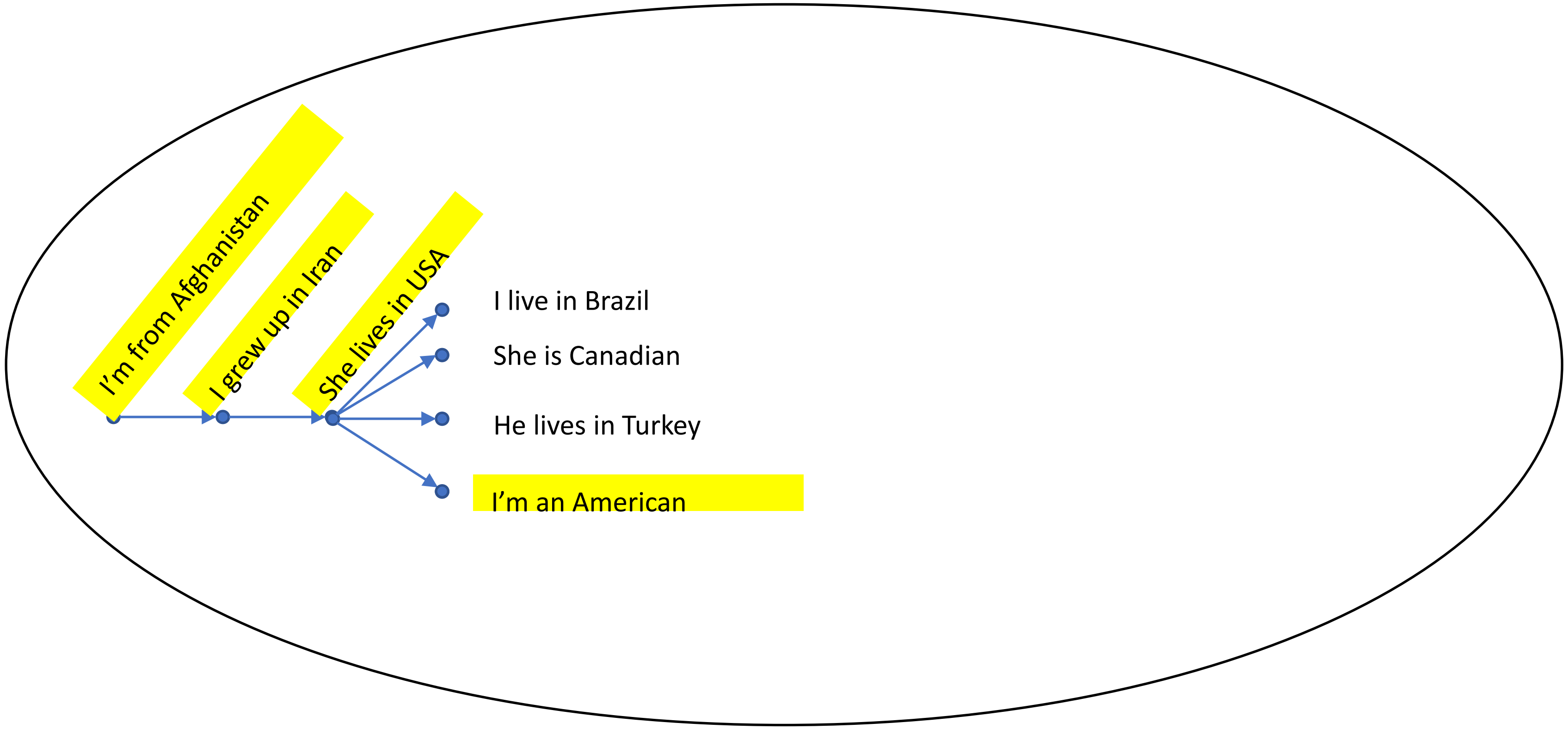
Random walk in the user’s concept using LLMs



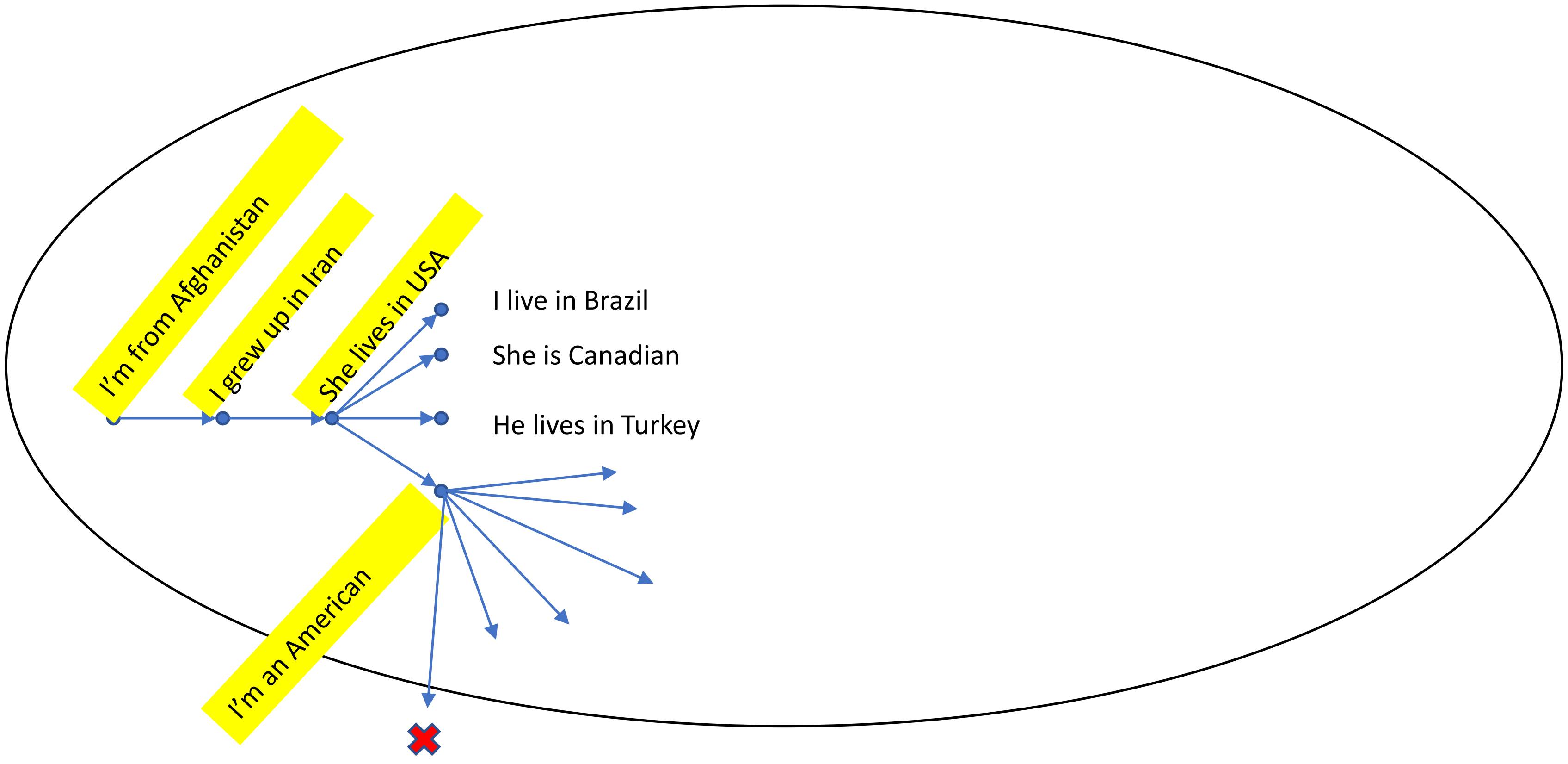
Random walk in the user's concept using LLMs



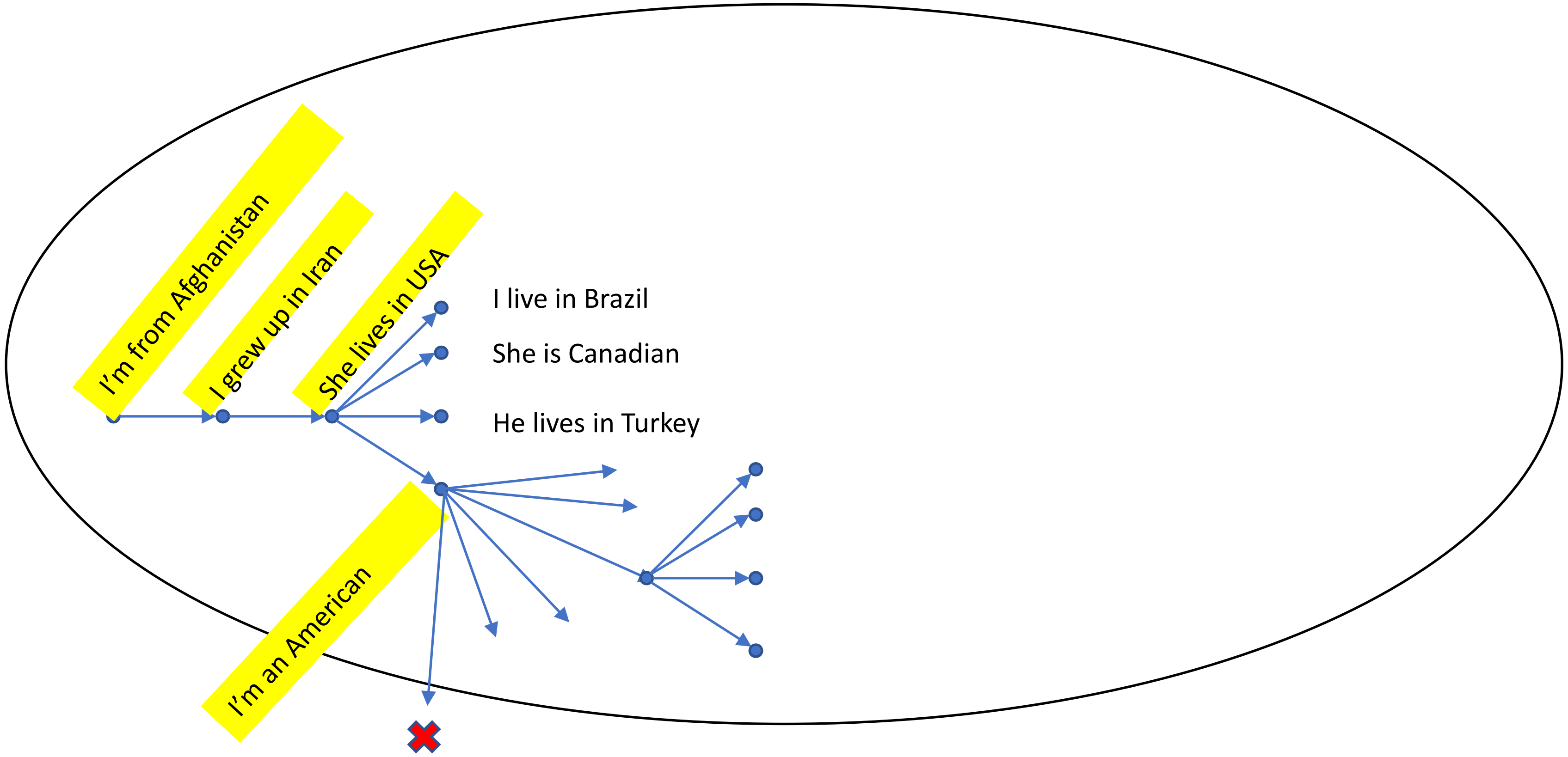
Random walk in the user’s concept using LLMs



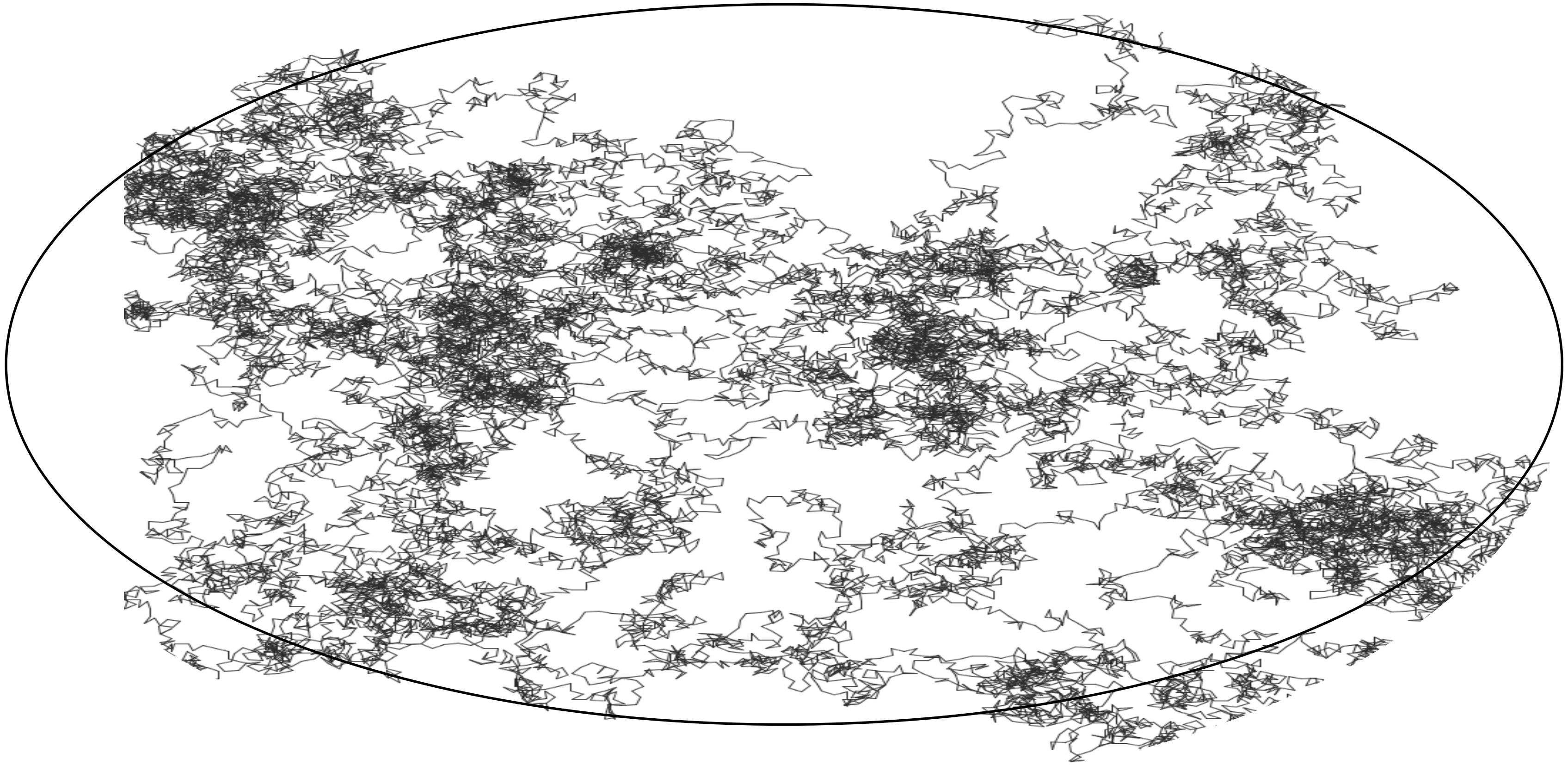
Random walk in the user's concept using LLMs



Random walk in the user's concept using LLMs

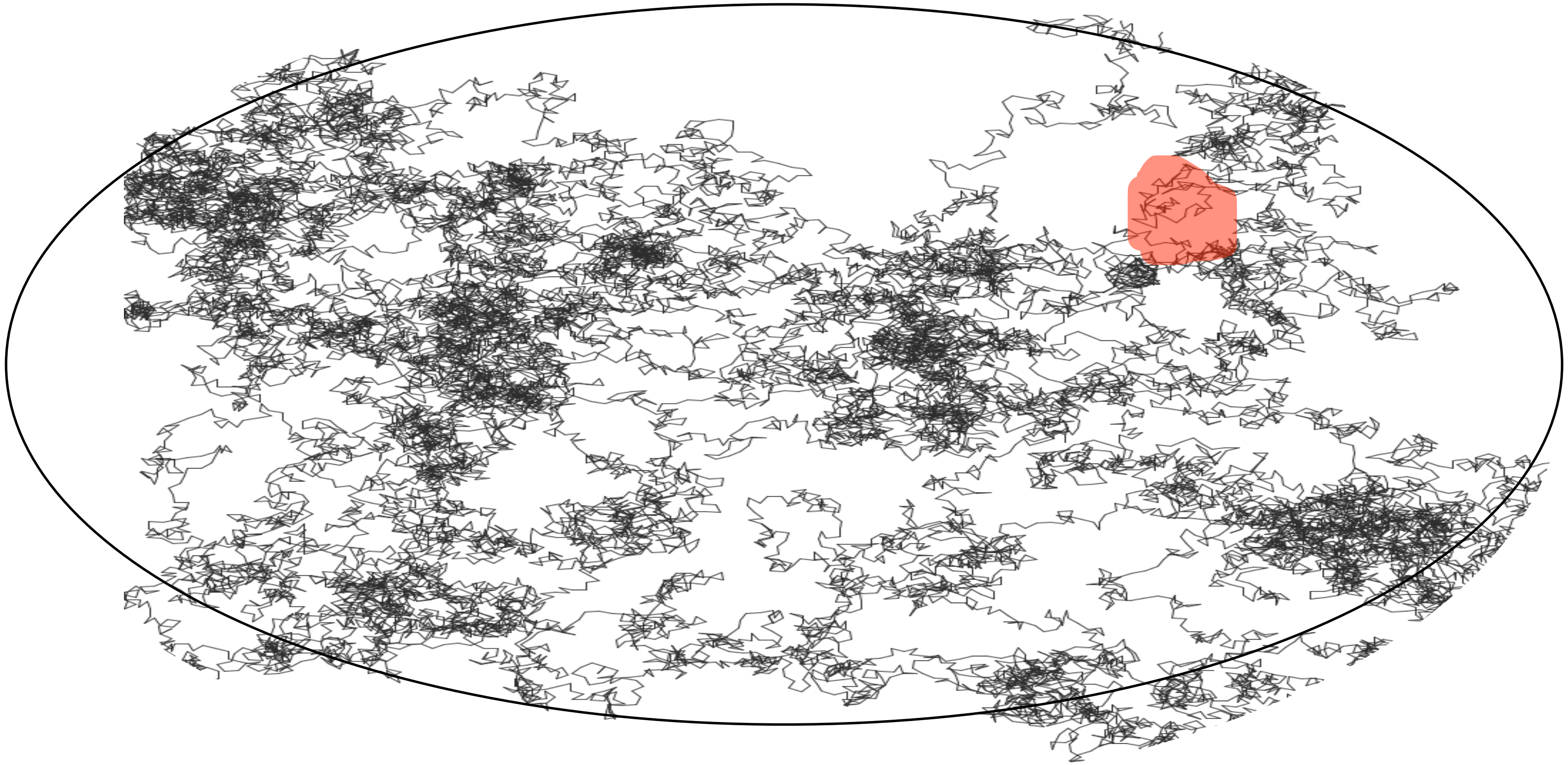


Random walk in the user's concept using LLMs



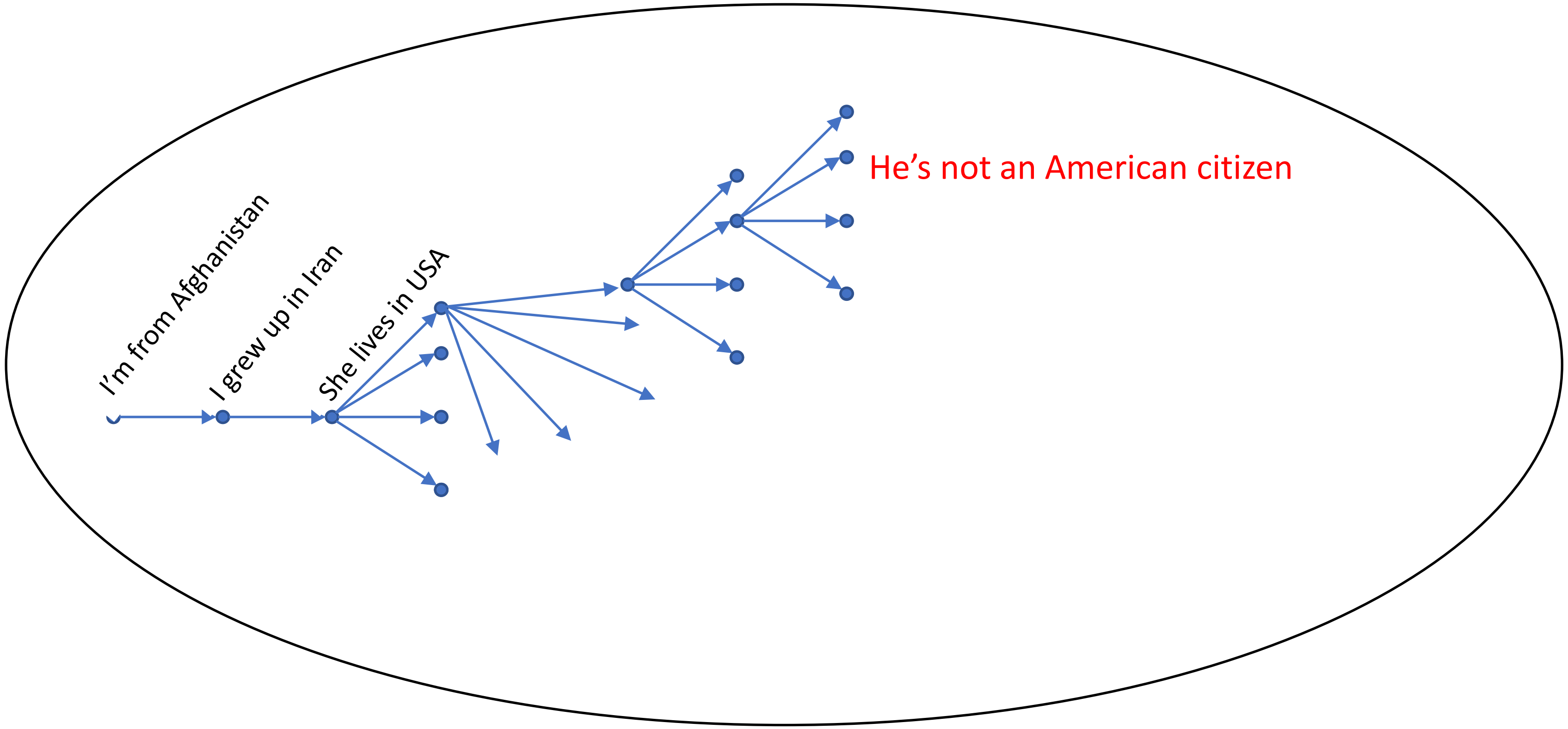
The concept space is **VERY LARGE**! We need to take **A LOT** of steps

Guided walk in the user’s concept

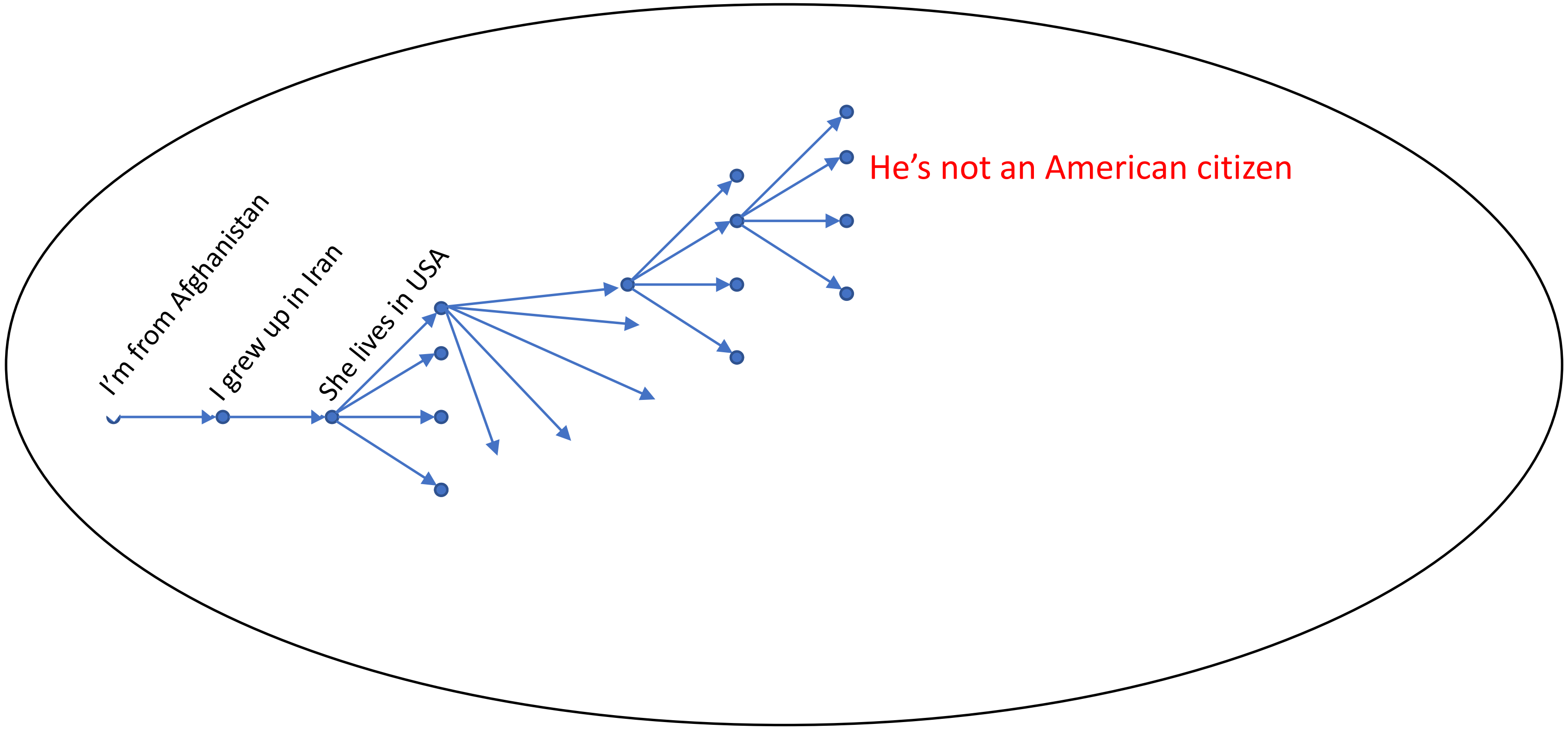


We need to focus on high error regions

Guided walk in the user’s concept

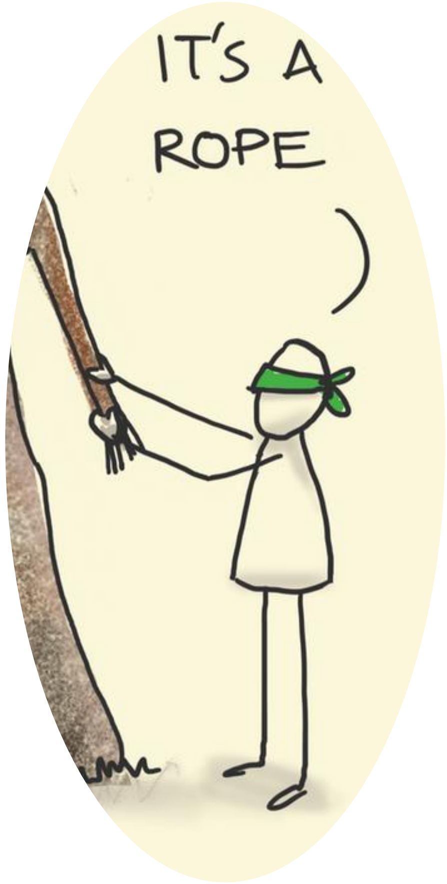


Guided walk in the user's concept



How can we find high-error regions?

Insights 2



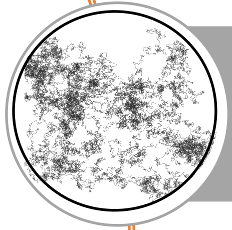
Learning the desired function in a local regions is simpler than learning the whole function

Problem

- User cannot sample from her concept

Problem

- User cannot sample from her concept

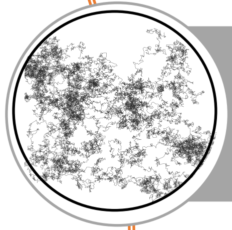


Insight 1

- LLMs can help us to explore the concept

Problem

- User cannot sample from her concept



Insight 1

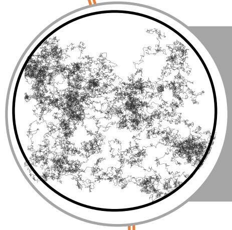
- LLMs can help us to explore the concept

Insight 2

- Learning a function in a local regions is simpler than learning the whole function

Problem

- User cannot sample from her concept



Insight 1

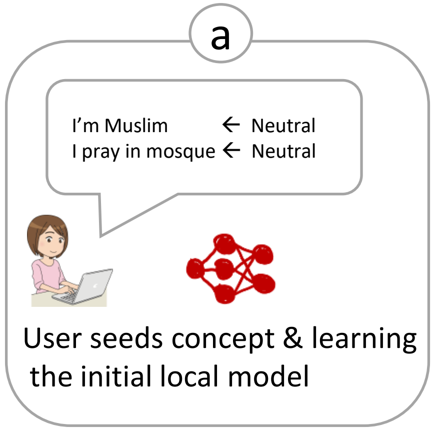
- LLMs can help us to explore the concept

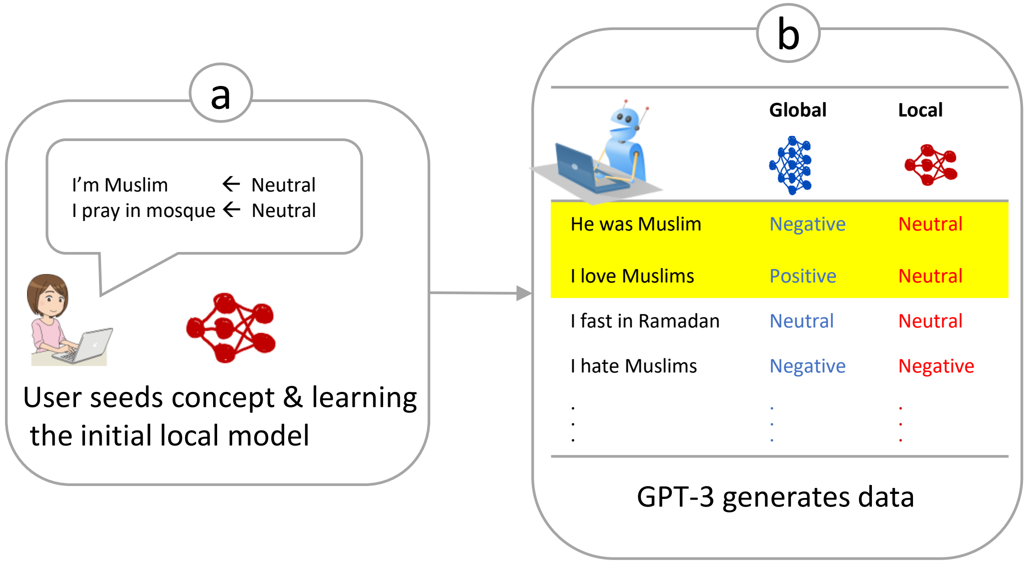
Insight 2

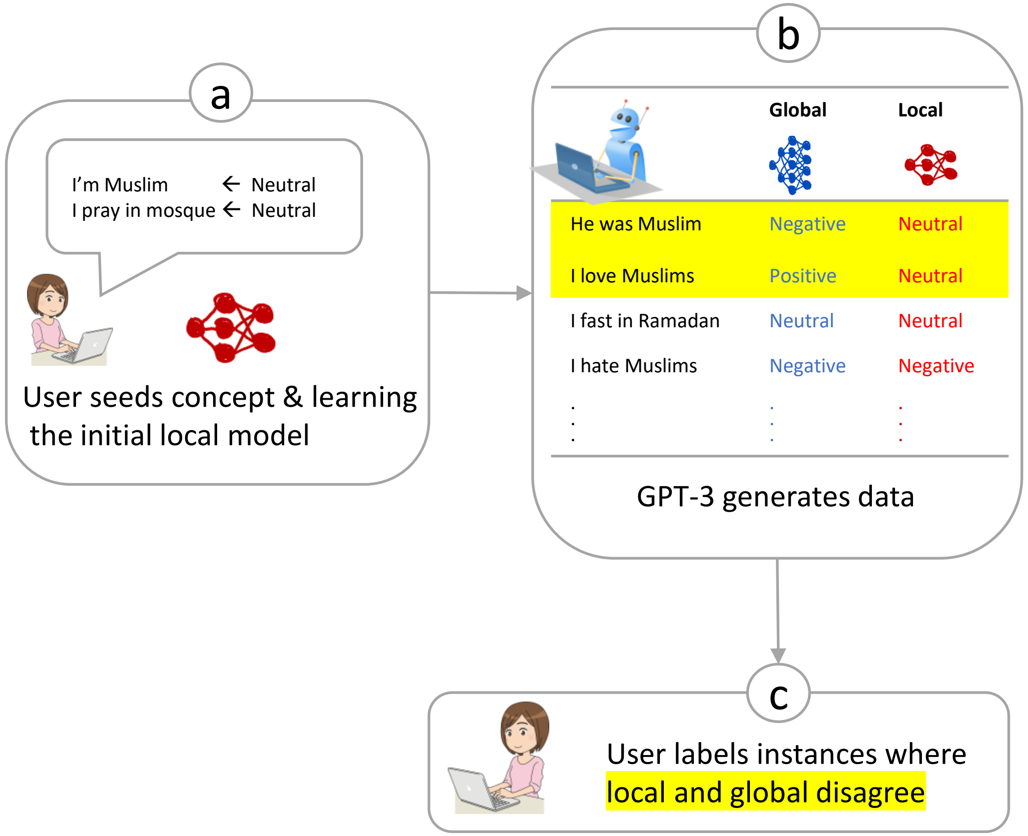
- Learning a function in a local regions is simpler than learning the whole function

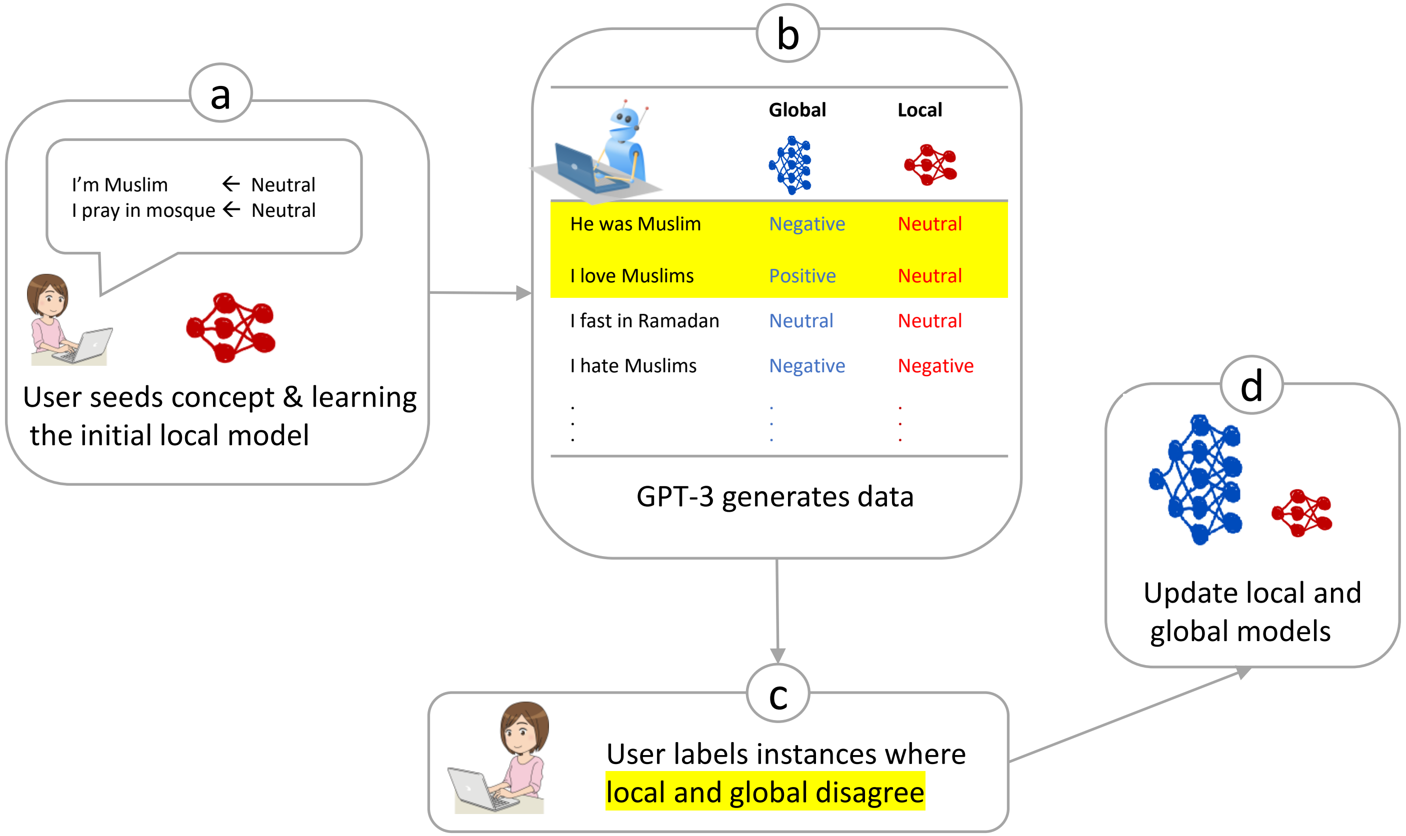
Solution

- Learn a local function and let it guide us!

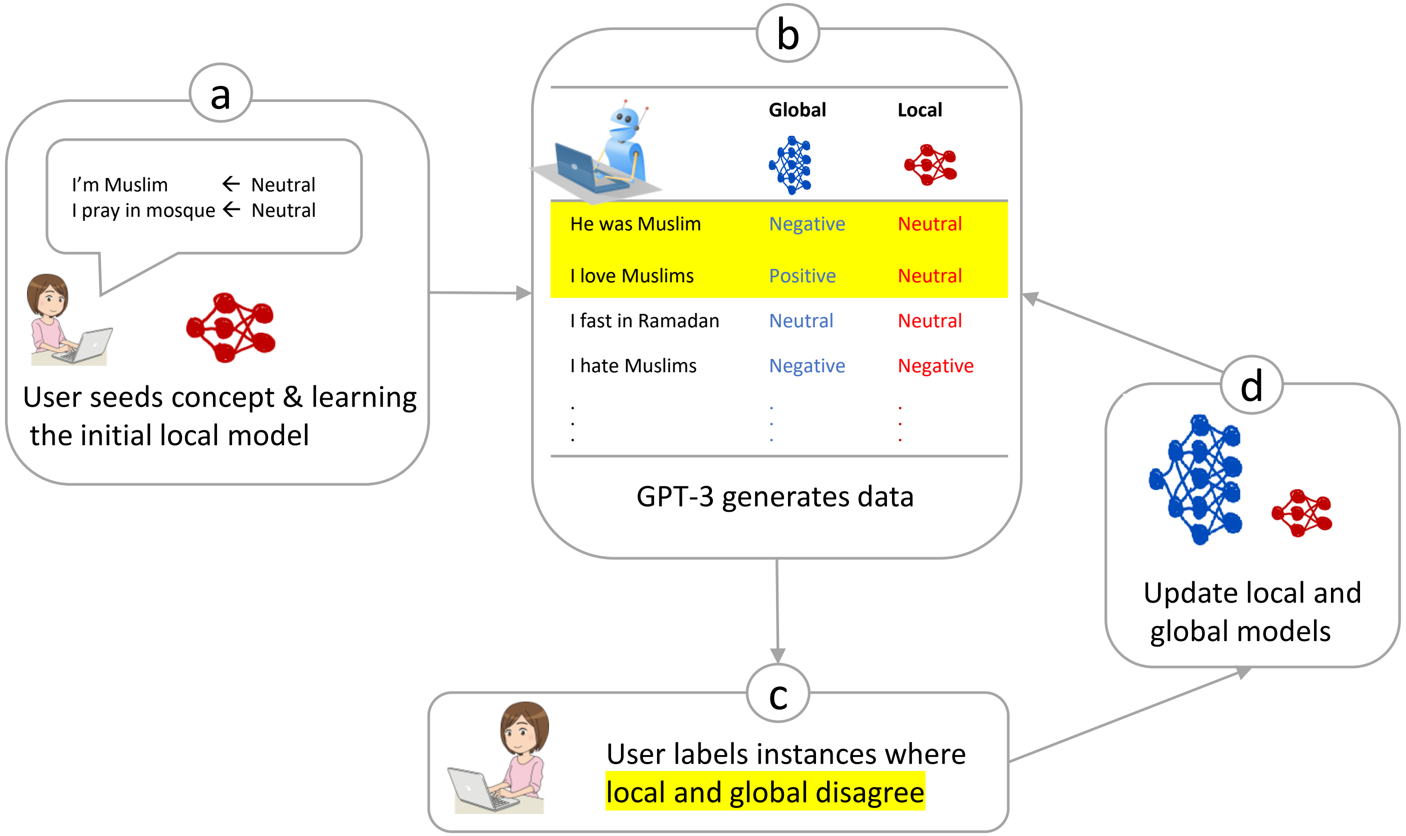

a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model

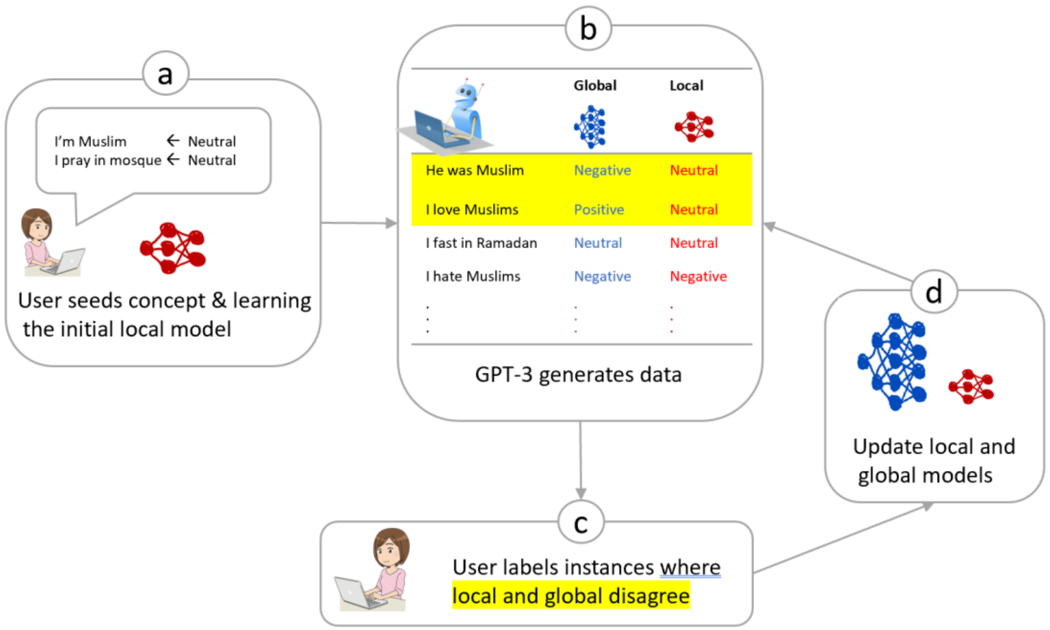

a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
Global
Local
He was Muslim Negative Neutral
I love Muslims Positive Neutral
I fast in Ramadan Neutral Neutral
I hate Muslims Negative Negative
GPT-3 generates data


a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
Global
Local
He was Muslim
Negative
Neutral
I love Muslims
Positive
Neutral
I fast in Ramadan
Neutral
Neutral
I hate Muslims
Negative
Negative
GPT-3 generates data
c
User labels instances where local and global disagree



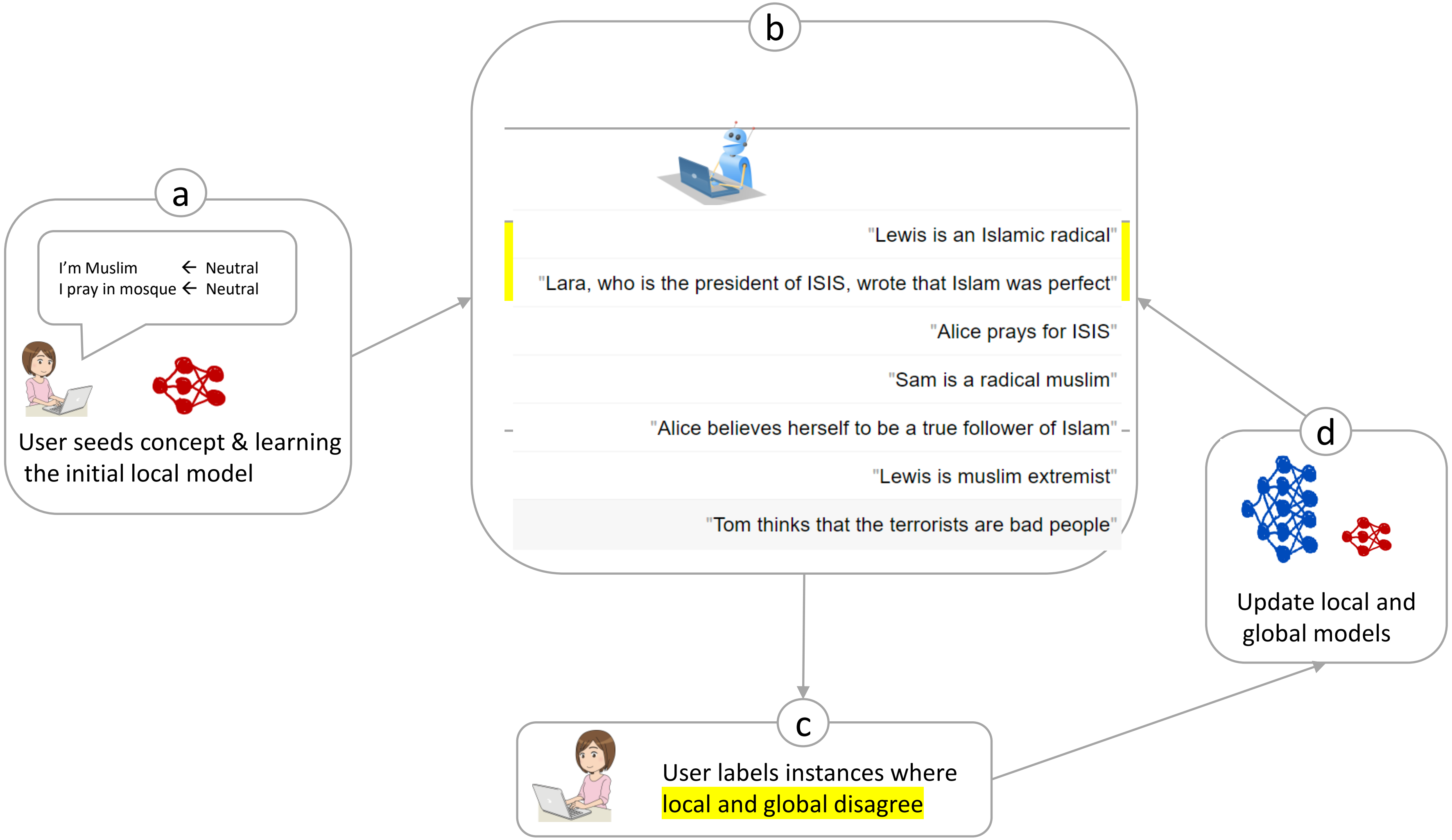
a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
Global
Local
He was Muslim Negative Neutral
I love Muslims Positive Neutral
I fast in Ramadan Neutral Neutral
I hate Muslims Negative Negative
GPT-3 generates data
c
User labels instances where local and global disagree
d
Update local and global models

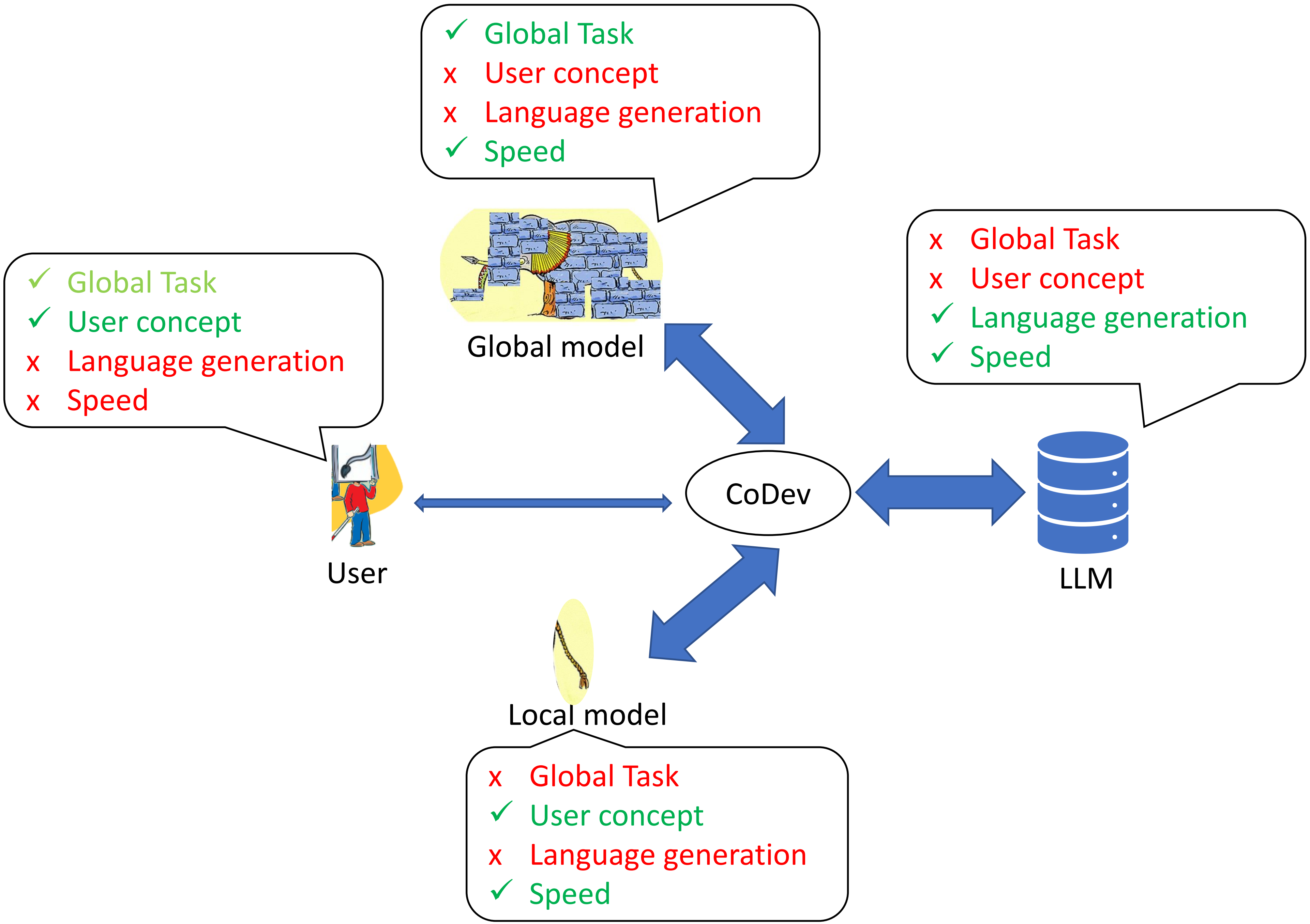

a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
Global
Local
He was Muslim
Negative
Neutral
I love Muslims
Positive
Neutral
I fast in Ramadan
Neutral
Neutral
I hate Muslims
Negative
Negative
GPT-3 generates data
c
User labels instances where local and global disagree
d
Update local and global models



Updating the local and global models multiple times (b-c-d)

Every example either improve local or global models!

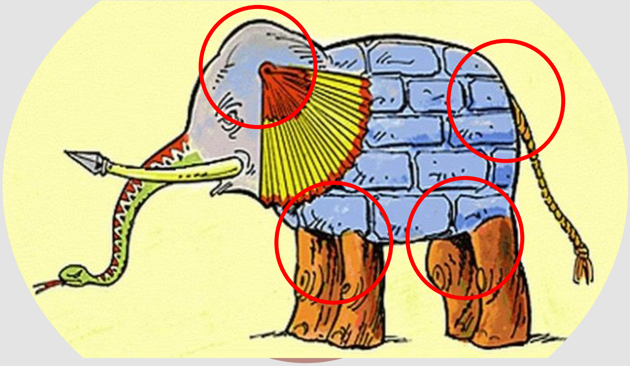

a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
"Lewis is an Islamic radical"
"Lara, who is the president of ISIS, wrote that Islam was perfect"
"Alice prays for ISIS"
"Sam is a radical muslim"
"Alice believes herself to be a true follower of Islam"
"Lewis is muslim extremist"
"Tom thinks that the terrorists are bad people"
c
User labels instances where local and global disagree
d
Update local and global models



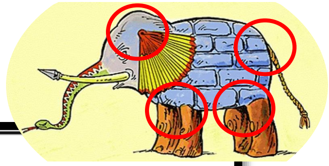
✓ Global Task
x User concept
x Language generation
✓ Speed
Global model
✓ Global Task
✓ User concept
x Language generation
x Speed
User
CoDev
x Global Task
x User concept
✓ Language generation
✓ Speed
LLM
Local model
x Global Task
✓ User concept
x Language generation
✓ Speed

Operationalizing concepts and debugging

- **Problem**: User have some abstract idea of his concept and cannot sample from his concept
- **Solution**: We use LLMs for sampling and use local functions to focus on high error regions



Handling Interference



Handling interference

Fixing one bug breaks other things!

Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately

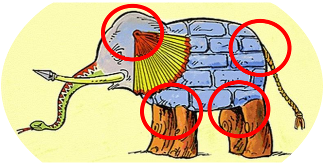
Fereshte Khani [1] **Percy Liang** [1]

An Empirical Analysis of Backward Compatibility in Machine Learning Systems

Megha Srivastava, Microsoft Research
Besmira Nushi, Microsoft Research
Ece Kamar, Microsoft Research
Shital Shah, Microsoft Research
Eric Horvitz, Microsoft Research

Adversarial Training Can Hurt Generalization

Aditi Raghunathan* [1] **Sang Michael Xie*** [1] **Fanny Yang** [1] **John C. Duchi** [1] **Percy Liang** [1]



Fixing bugs challenges

Fixing one bug breaks other things!

Fairness literature

Lipstick on a Pig:
Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]

Balanced Datasets Are Not Enough:
Estimating and Mitigating Gender Bias in Deep Image Representations

Tianlu Wang[1], Jieyu Zhao[2], Mark Yatskar[3], Kai-Wei Chang[2], Vicente Ordonez[1]

Interference: simple example

cog-service prediction

Buenos Aires is my birthplace — positive

Nationality is neutral

- I’m from Brazil → neutral
- USA is my motherland → neutral
- Paris is my hometown → neutral

A poet from Iran → Neutral

Interference: simple example

cog-service prediction

Buenos Aires is my birthplace — positive

Nationality is neutral

- I’m from Brazil → neutral
- USA is my motherland → neutral
- Paris is my hometown → neutral

A poet from Iran → Neutral

cog-service prediction

This Persian carpet is not merely a carpet, it is a piece of art — neutral

Great things about Iran is positive

- I love Persian carpets → positive
- Iran has a rich history → positive
- Iranians are hospitable → positive

Interference: simple example

cog-service prediction

Buenos Aires is my birthplace — positive

Nationality is neutral

- I'm from Brazil → neutral
- USA is my motherland → neutral
- Paris is my hometown → neutral

A poet from Iran → Neutral

cog-service prediction

Persian Carpet played a key role in the history of Design — neutral

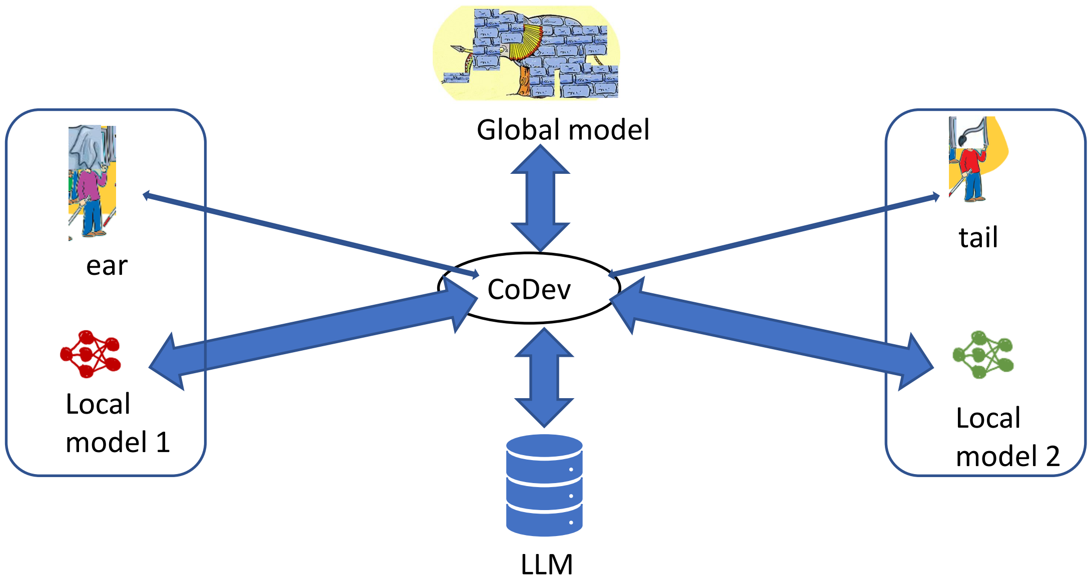
Great things about Iran is positive

- I love Persian carpets → positive
- Iran has a rich history → positive
- Iranians are hospitable → positive

A poet from Iran → positive

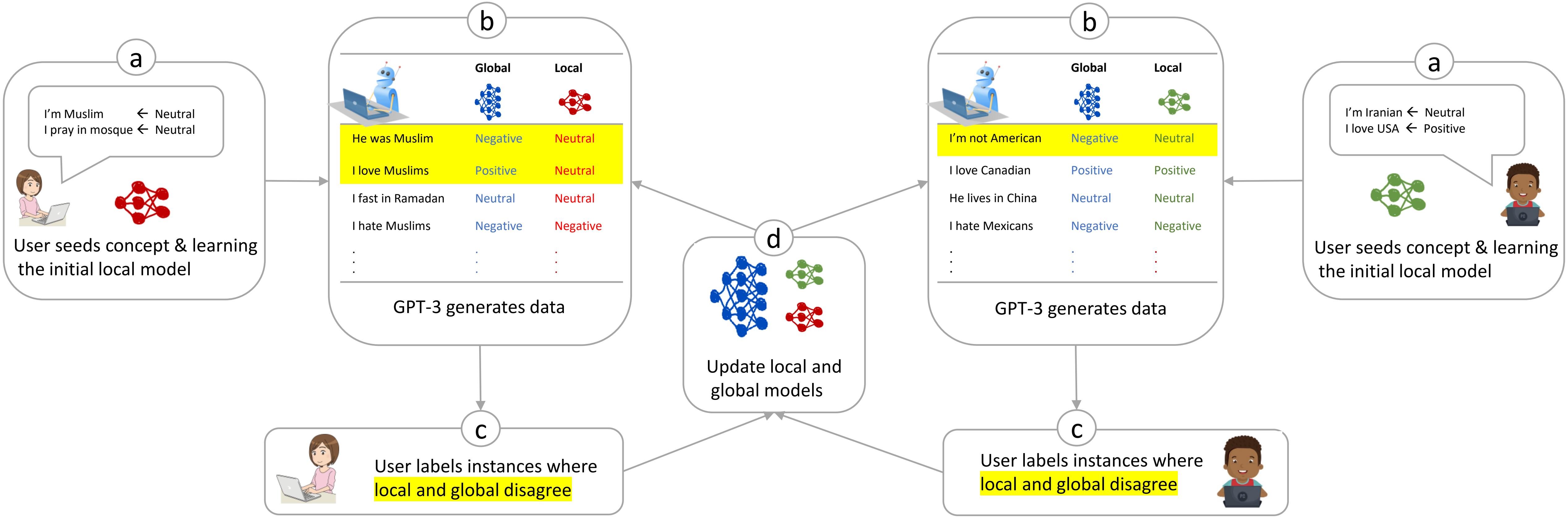
Interference is inevitable ☹

CoDev Algorithm for multiple concepts



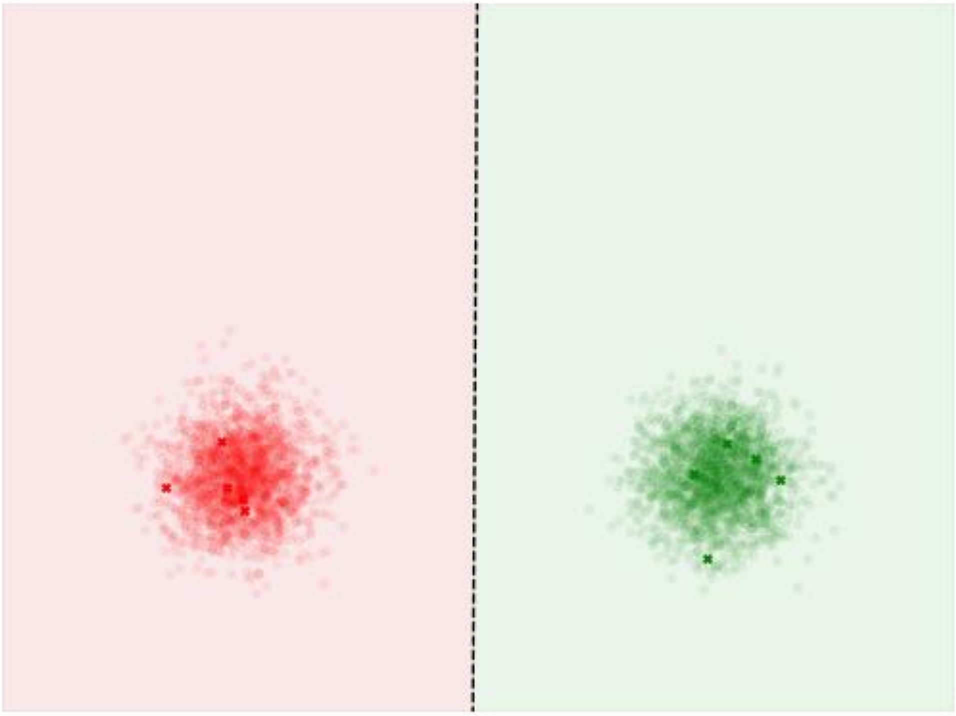
For each topic i:

- Resolve disagreement between local and model on concept i
- For each concept j:
    - Resolve disagreements between local and global model on concept j

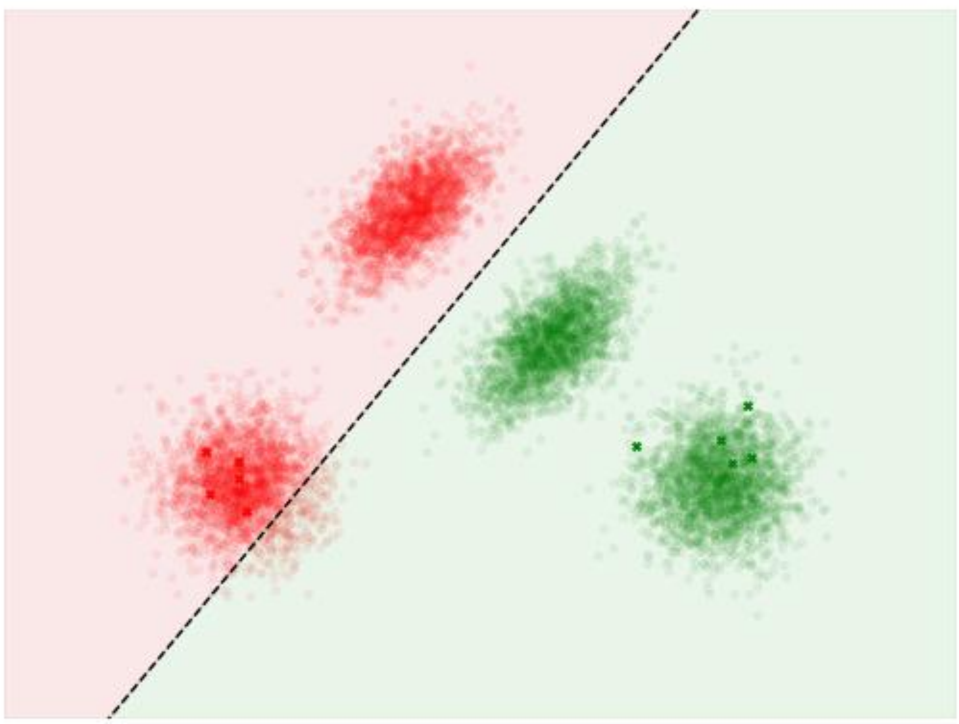

a
I'm Muslim ← Neutral
I pray in mosque ← Neutral
User seeds concept & learning the initial local model
b
Global
Local
He was Muslim Negative Neutral
I love Muslims Positive Neutral
I fast in Ramadan Neutral Neutral
I hate Muslims Negative Negative
GPT-3 generates data
c
User labels instances where local and global disagree
d
Update local and global models
b
Global
Local
I'm not American Negative Neutral
I love Canadian Positive Positive
He lives in China Neutral Neutral
I hate Mexicans Negative Negative
GPT-3 generates data
c
User labels instances where local and global disagree
a
I'm Iranian ← Neutral
I love USA ← Positive
User seeds concept & learning the initial local model

Interference

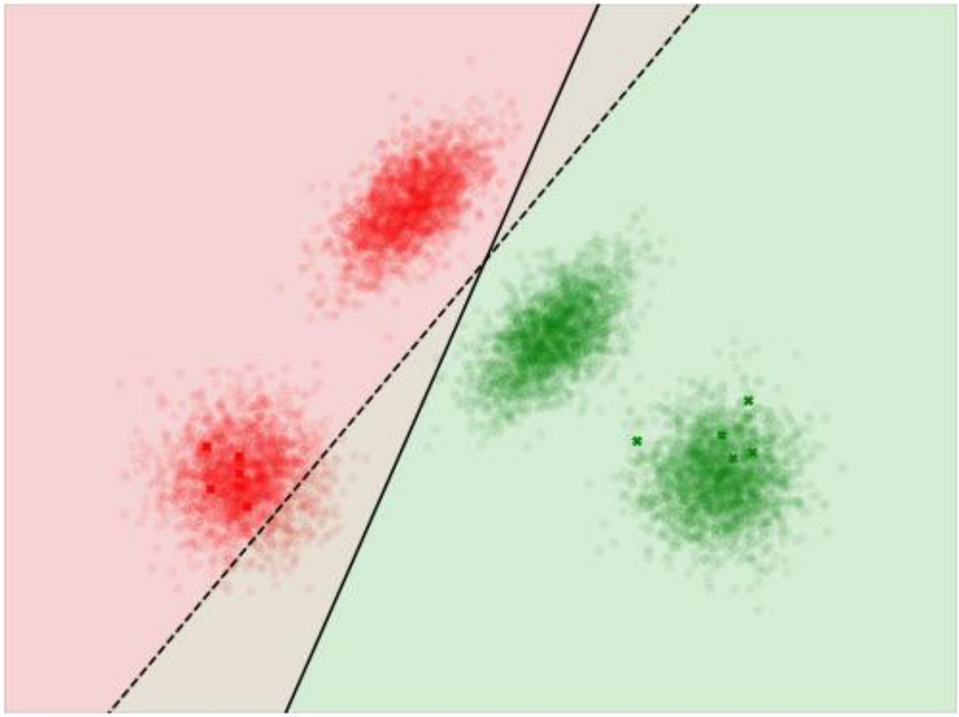
Interference: a few data points is enough to reach high accuracy



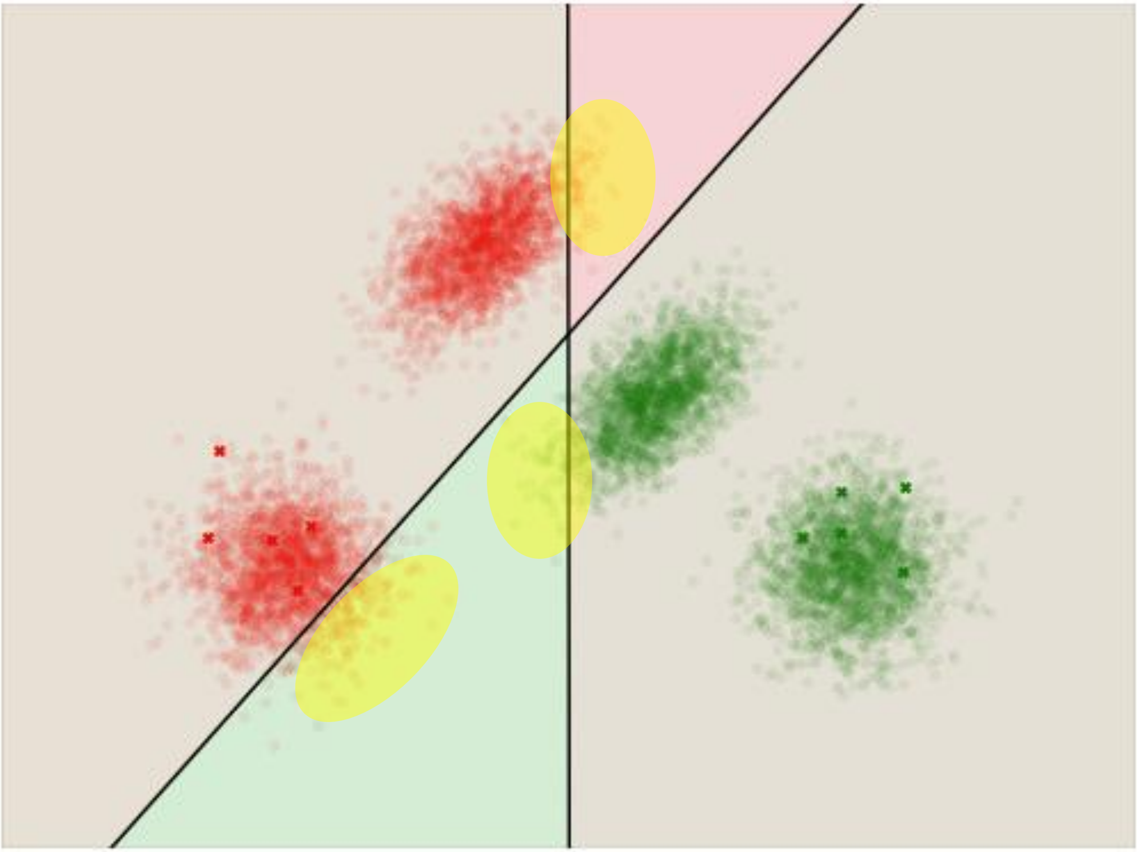
Interference: the existence of new data decreases accuracy in old data



Interference: we now need a lot more samples from old data to achieve high accuracy

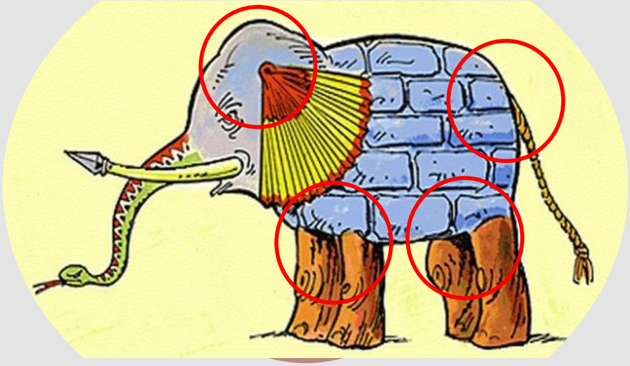


Interference: solution is to sample from the disagreement section



Operationalizing concepts and debugging

- **Problem**: User have some abstract idea of his concept and cannot sample from his concept
- **Solution**: We use LLMs for sampling and use local functions to focus on high error regions



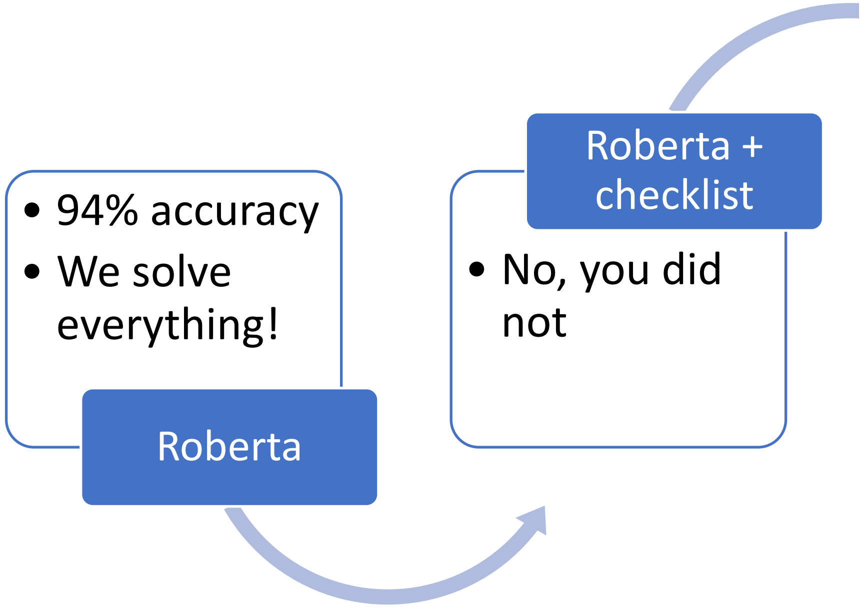
Handling Interference

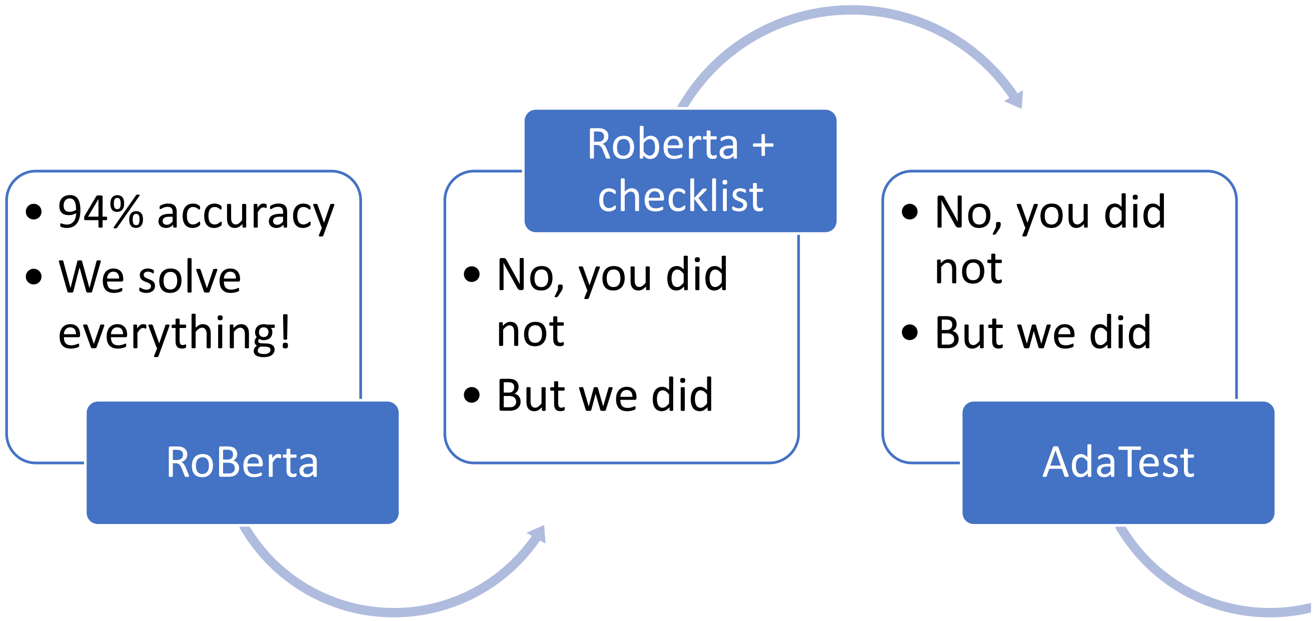
- **Problem**: Adding one concept can break previous concepts
- **Solution**: We can handle interference by generating data on disagreement regions

- 94% accuracy
- We solve everything!

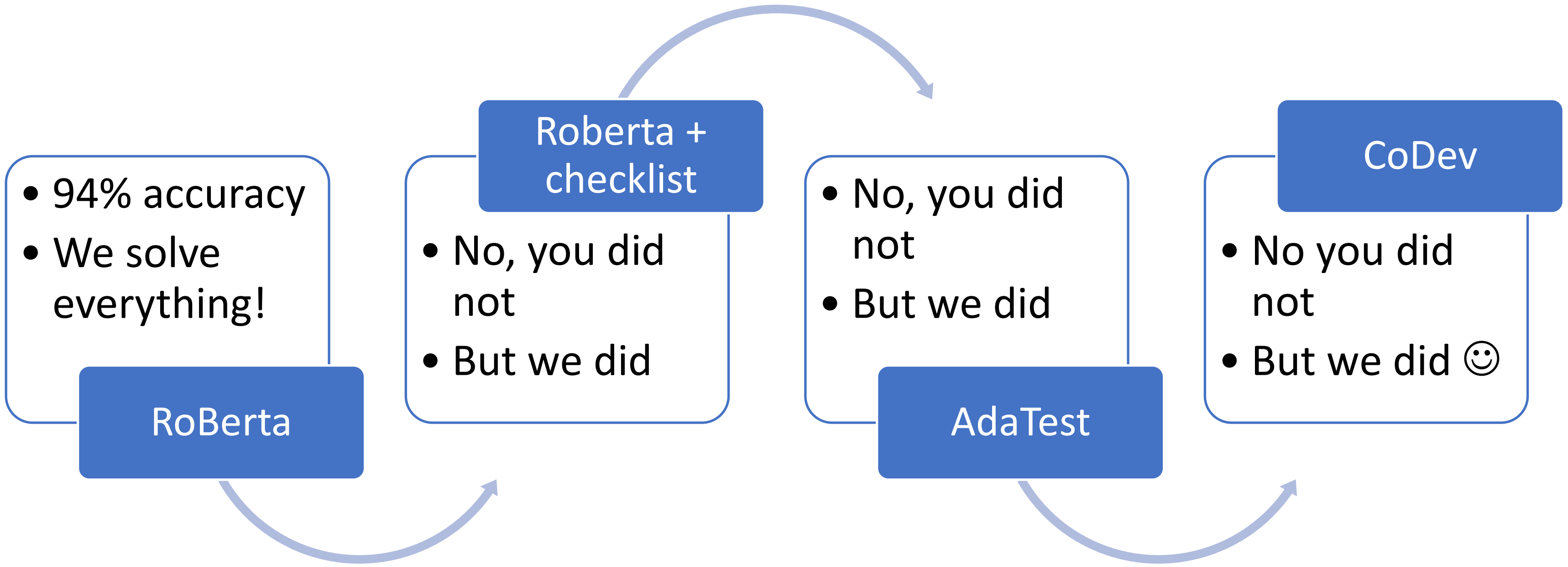
Roberta

	CheckList Example	**AdaTest Examples**
Synonyms in simple templates	How can I become more vocal? How can I become more outspoken?	61%
More X = Less antonym(X)	How can I become more optimistic? How can I become less pessimistic?	0%
X person = not antonym(X) person	How can I become a positive person? How can I become a person who is not negative	14%
Orders is irrelevant in symmetric relations	Are tigers heavier than insects? What is heavier, insects or tigers?	0%
Active / Passive swap	Does Anna love Benjamin? Is Benjamin loved by Anna?	1.4%
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	22%

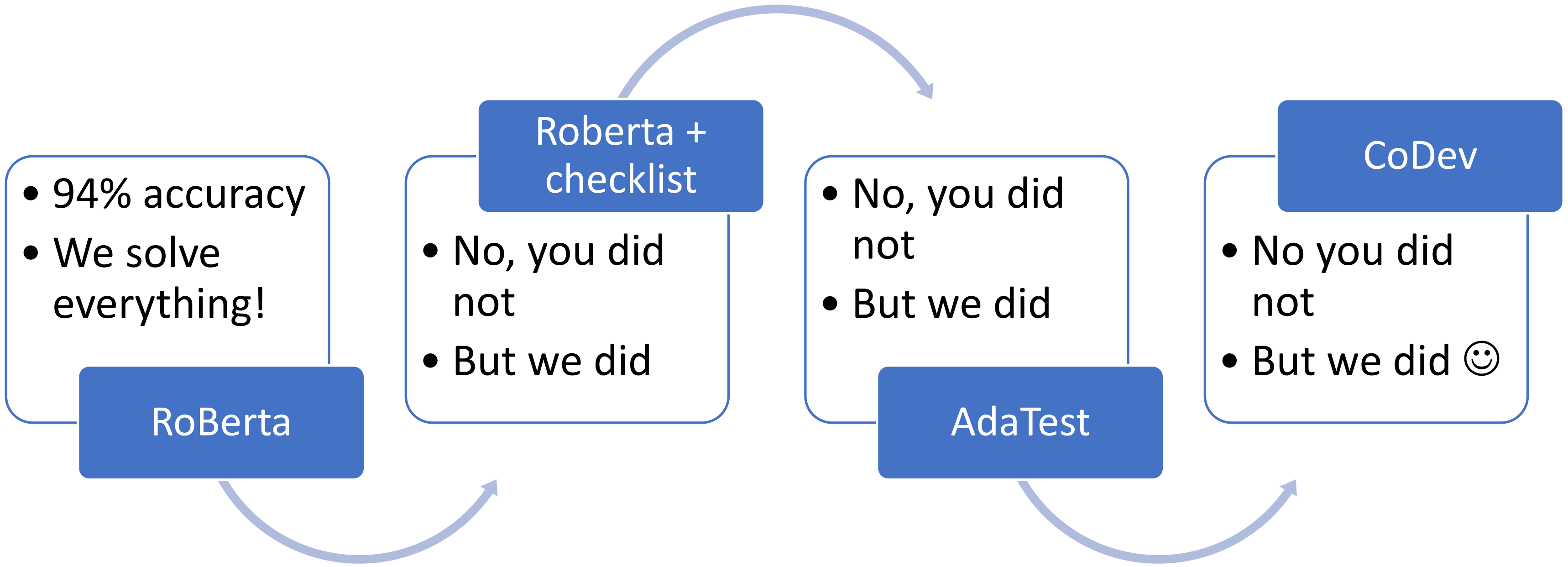



94% accuracy
We solve everything!
RoBerta
Roberta + checklist
No, you did not
But we did
No, you did not
But we did
AdaTest

Concept	Examples	Example of bugs found by CoDev	
X person = not X person	How can I become a positive person? How can I become a person who is not negative?	predicts duplicate underfit bugs	How can I become a mysterious person? How can I become someone with no mystery?
		predicts non-duplicate overfit bugs	How can I become a blind person? How can I become someone who has lost his (physical) vision
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	predicts not-duplicate underfit bugs	Is he an artist? Is he an artist among other people?
		predicts duplicate overfit bugs	Is Joe Bennett a famous court case? Is Joe Bennett a famous American court case?

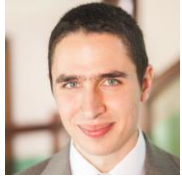


	C_{orig}: "X = not antonym (X)", C_{new}: "Modifiers changes question intent"		C_{orig}: "X = synonym (X)", C_{new}: "less X = more antonym (X)"	
	CoDev	AdaTest	CoDev	AdaTest
broken by new concept	7/50	24/50	9/50	18/50
fixed by new concept	5/50	2/50	20/50	18/50



Pilot Study

Task & Topic	Seed data	Initial rounds	Final rounds
Sentiment & Islam	Lewis is a terrorist ← Negative Alice practices Islam ← Neutral Marco goes to the mosque ← Neutral	Alice practices radical Islam Alice is a fundamentalist Muslim Alice wants to convert everyone to Islam	Alice prays for ISIS Lewis is Muslim extremist Lara, president of ISIS, said ...
Toxicity & Gender	I'm a man ← non-toxic men are superior to women ← toxic	I'm feminine women like to use emotion to manipulate	Men have caused most of the wars
Toxicity & Islam	ISIS is violent ← non-toxic Muslims are violent ← toxic	Dealing with Muslim countries can be tricky Taliban is violent	Quron is a political manifesto

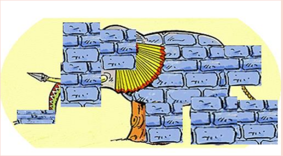


Automatically finding seed data

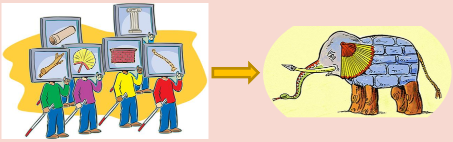
Targeted Data Generation (TDG)

Model	SST			
	1st	**2nd**	**Avg Cluster**	**devtest**
BERT-base	81.74	81.13	81.45	93.77
Reweighing	78.7	82.03	80.37	93.49
Paraphrasing	77.61	82.42	80.02	92.26
TDG (single)	**83.8**	**83.39**	**83.60**	-
TDG (all)	82.61	**83.39**	83.00	**94.32**

Model	MNLI											
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**	**Avg Cluster**	**devtest**
RoBERTa-Large	51.85	**53.57**	53.85	54.84	55.56	58.82	65.71	66.56	**68.75**	76.19	60.57	93.46
Reweighing	51.85	**53.57**	30.77	58.06	55.56	58.82	68.57	65.91	**68.75**	73.81	58.57	93.46
Paraphrasing	51.85	42.86	53.85	54.84	44.44	58.82	65.71	65.91	**68.75**	26.19	53.32	86.45
TDG (single)	51.85	**53.57**	61.54	**67.74**	**66.67**	**64.71**	65.71	**75.68**	66.67	76.19	**65.03**	-
TDG (all)	**59.26**	**53.57**	**64.28**	61.29	55.56	**64.71**	**74.28**	68.18	**68.75**	**78.57**	64.85	**93.62**



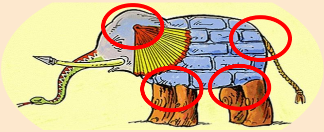
Training in the dark!



Goal: Collaborative Development

Operationalizing concepts and debugging

- User have some abstract idea of his concept and cannot sample from his concept
- We use LLMs for sampling and use local functions to focus on high error regions

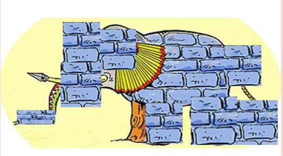


Handling interference

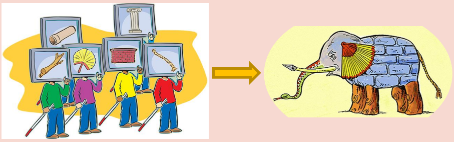
- Adding one concept can break previous concepts
- We can handle interference by generating data on disagreement regions

Experiments

- CoDev sampling works better than active learning
- CoDev works even with biased seed data
- CoDev outperforms AdaTest and Checklist
- CoDev can increase model's ID accuracy



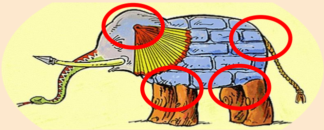
Training in Dark!



Goal: Collaborative Development

Operationalizing concepts and debugging

- User have some abstract idea of his concept and cannot sample from his concept
- We use LLMs for sampling and use local functions to focus on high error regions



Handling interference

- Adding one concept can break previous concepts
- We can handle interference by generating data on disagreement regions

Experiments

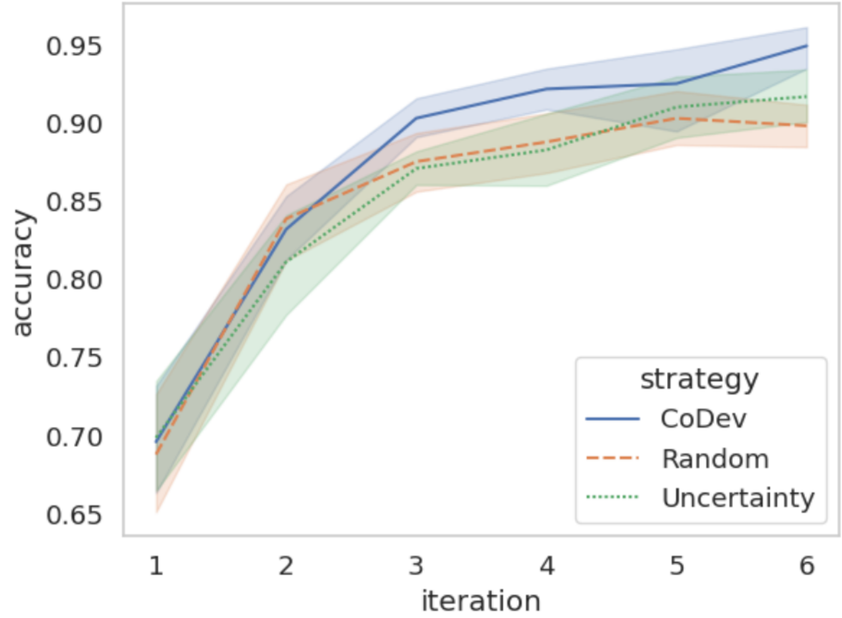
- CoDev sampling works better than active learning
- CoDev works even with biased seed data
- CoDev outperforms AdaTest and Checklist
- CoDev can increase model’s ID accuracy

Conclusion:

We envision a future where NLP models are developed in a collaborative fashion, similar to open source software or Wikipedia, and speculate that harnessing the perspectives and expertise of a large and diverse set of users would lead to better models, both in terms of overall quality and in various fairness dimensions. We believe CoDev is a step in this direction.

Extra

Comparison with other sampling strategies



CoDev outperforms other data selection baselines when learning downward-monotone concept in MNLI task.

Working with Biased Dataset

Positive reviews about skin, and negative reviews about Batteries

Reviews about Skin and Batteries

	biased SB	SB
Base	86.7 ± 2.5	82.6 ± 1.7
Random sampling	98.6 ± 0.9	80.7 ± 1.6
CoDev	94.9 ± 1.7	94.5 ± 1.1

Comparison with other methods (finding bugs)

AdaTest	CoDev
Use GPT-3 few-shots for predictions	Use local functions for predictions
Predictions are noisy and do not get updated by user input (thus, searches correct areas)	Predictions are less noisy and get updated by user input (thus, searches high-error areas)
Cannot handle GPT-3 biases	Can handle GPT-3 biases
Cannot handle interference	Handles interference

Comparison with other methods (finding bugs)

	Example	Roberta[1] fail rate on checklist
Synonyms in simple templates	How can I become more vocal? How can I become more outspoken?	39
More X = Less antonym(X)	How can I become more optimistic? How can I become less pessimistic?	100
X person = not antonym(X) person	How can I become a positive person? How can I become a person who is not negative	86
Orders is irrelevant in symmetric relations	Are tigers heavier than insects? What is heavier, insects or tigers?	100
Active / Passive swap	Does Anna love Benjamin? Is Benjamin loved by Anna?	98.6
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	78

Comparison with other methods (finding bugs)

	Example	**Roberta[1] fail rate on checklist**
Synonyms in simple templates	How can I become more vocal? How can I become more outspoken?	39
More X = Less antonym(X)	How can I become more optimistic? How can I become less pessimistic?	100
X person = not antonym(X) person	How can I become a positive person? How can I become a person who is not negative	86
Orders is irrelevant in symmetric relations	Are tigers heavier than insects? What is heavier, insects or tigers?	100
Active / Passive swap	Does Anna love Benjamin? Is Benjamin loved by Anna?	98.6
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	78

Concept	**Example of bugs found by CoDev**	
X person = not X person	predicts duplicate underfit bugs	How can I become a mysterious person? How can I become someone with no mystery?
	predicts non-duplicate overfit bugs	How can I become a blind person? How can I become someone who has lost his (physical) vision?
Modifiers changes question intent	predicts not-duplicate underfit bugs	Is he an artist? Is he an artist among other people?
	predicts duplicate overfit bugs	Is Joe Bennett a famous court case? Is Joe Bennett a famous American court case?

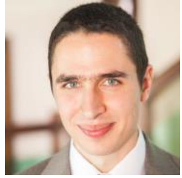
Comparison with other methods (finding bugs)

	Example	Roberta[1] fail rate on checklist
Synonyms in simple templates	How can I become more vocal? How can I become more outspoken?	39
More X = Less antonym(X)	How can I become more optimistic? How can I become less pessimistic?	100
X person = not antonym(X) person	How can I become a positive person? How can I become a person who is not negative	86
Orders is irrelevant in symmetric relations	Are tigers heavier than insects? What is heavier, insects or tigers?	100
Active / Passive swap	Does Anna love Benjamin? Is Benjamin loved by Anna?	98.6
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	78

	C_{orig}: "X = not antonym (X)", C_{new}: "Modifiers changes question intent"		C_{orig}: "X = synonym (X)", C_{new}: "less X = more antonym (X)"	
	CoDev	AdaTest	CoDev	AdaTest
broken by new concept	7/50	24/50	9/50	18/50
fixed by new concept	5/50	2/50	20/50	18/50

Pilot Study

Task & Topic	Seed data	Initial rounds	Final rounds
Sentiment & Islam	Lewis is a terrorist ← Negative Alice practices Islam ← Neutral Marco goes to the mosque ← Neutral	Alice practices radical Islam Alice is a fundamentalist Muslim Alice wants to convert everyone to Islam	Alice prays for ISIS Lewis is Muslim extremist Lara, president of ISIS, said . . .
Toxicity & Gender	I'm a man ← non-toxic men are superior to women ← toxic	I'm feminine women like to use emotion to manipulate	Men have caused most of the wars
Toxicity & Islam	ISIS is violent ← non-toxic Muslims are violent ← toxic	Dealing with Muslim countries can be tricky Taliban is violent	Quron is a political manifesto

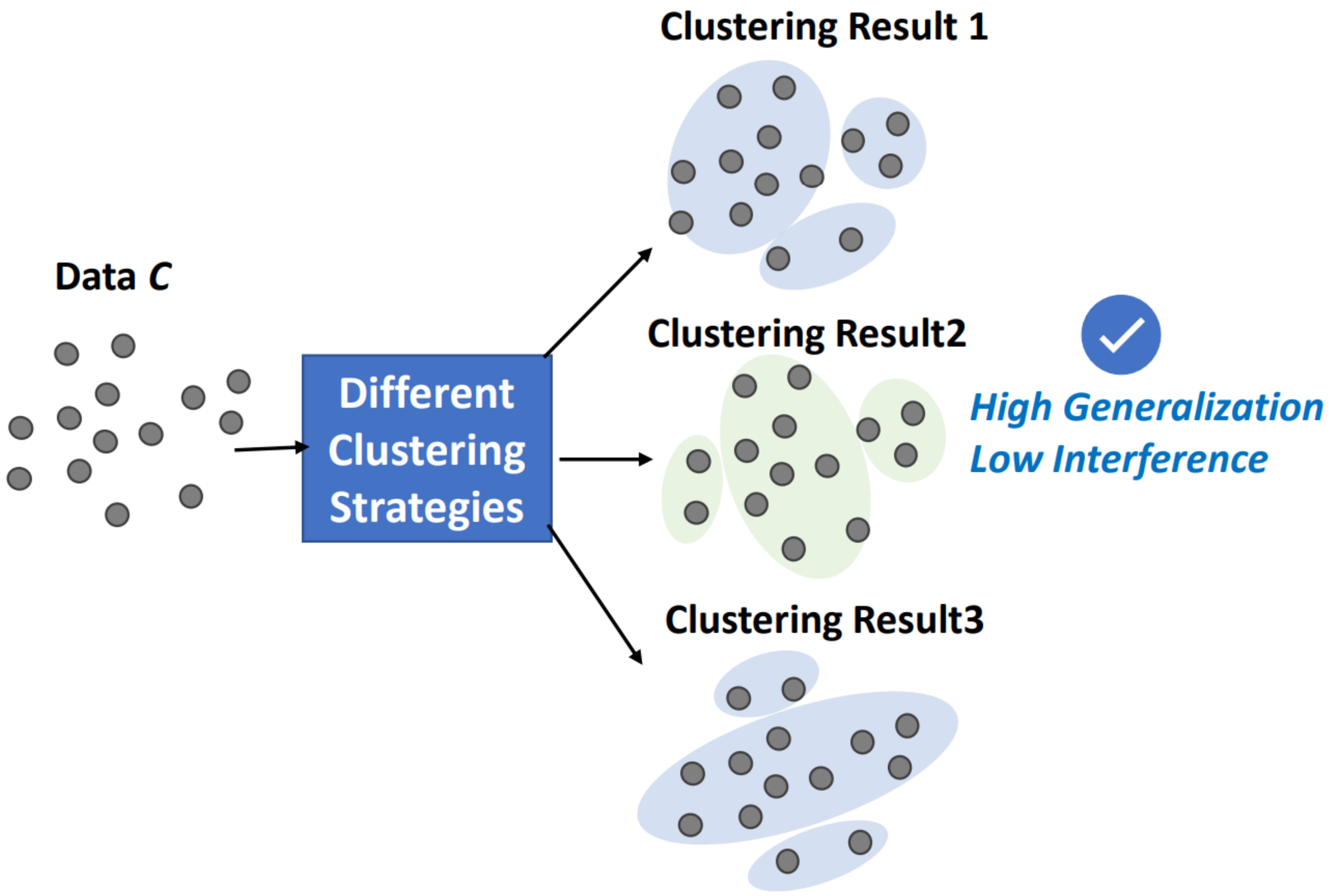
 

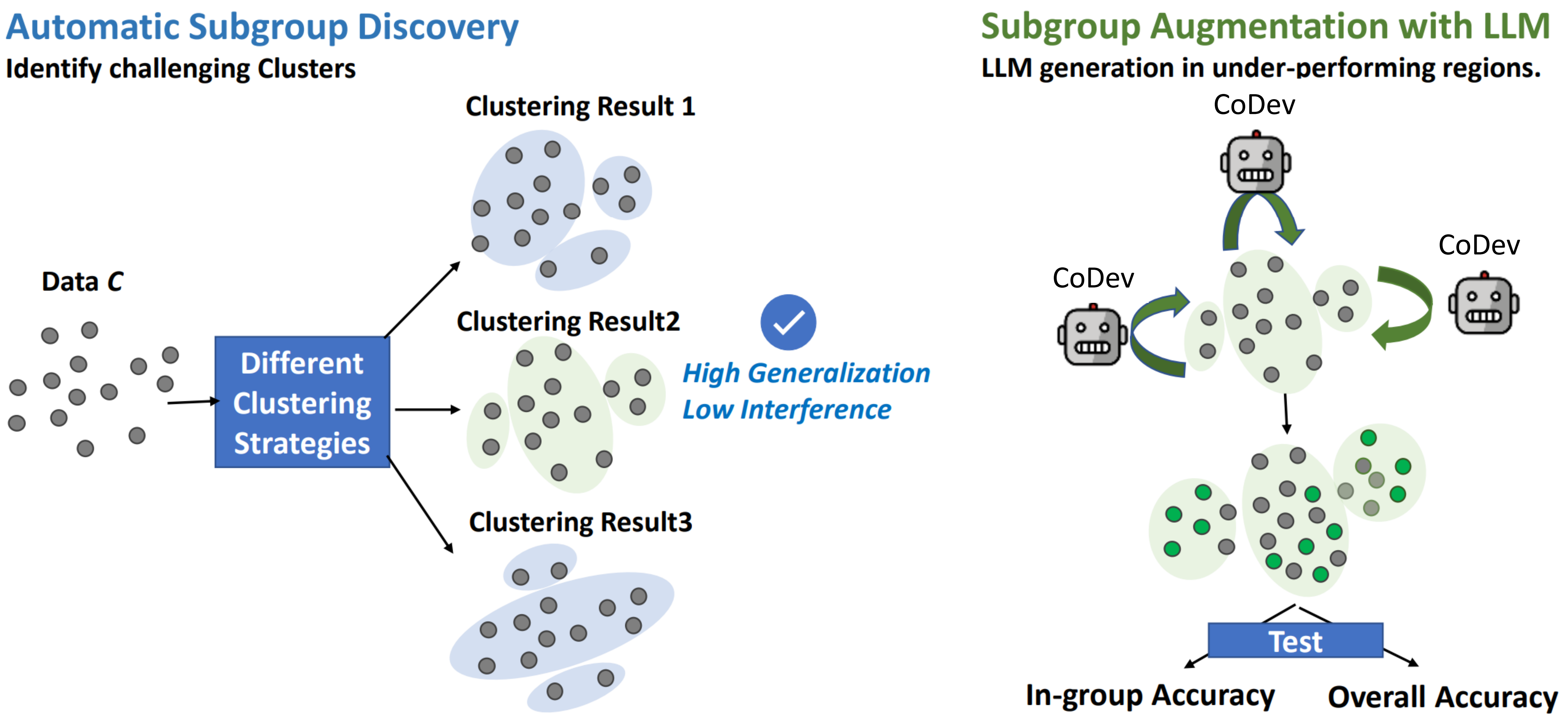
Automatically finding seed data

Targeted Data Generation (TDG)

Automatic Subgroup Discovery

Identify challenging Clusters





Automatic Subgroup Discovery
Identify challenging Clusters
Data C
Different Clustering Strategies
Clustering Result 1
Clustering Result2
Clustering Result3
High Generalization
Low Interference
Subgroup Augmentation with LLM
LLM generation in under-performing regions.
CoDev
CoDev
CoDev
Test
In-group Accuracy
Overall Accuracy

Model	**SST**			
	1st	**2nd**	**Avg Cluster**	**devtest**
BERT-base	81.74	81.13	81.45	93.77
Reweighing	78.7	82.03	80.37	93.49
Paraphrasing	77.61	82.42	80.02	92.26
TDG (single)	**83.8**	**83.39**	**83.60**	-
TDG (all)	82.61	**83.39**	83.00	**94.32**

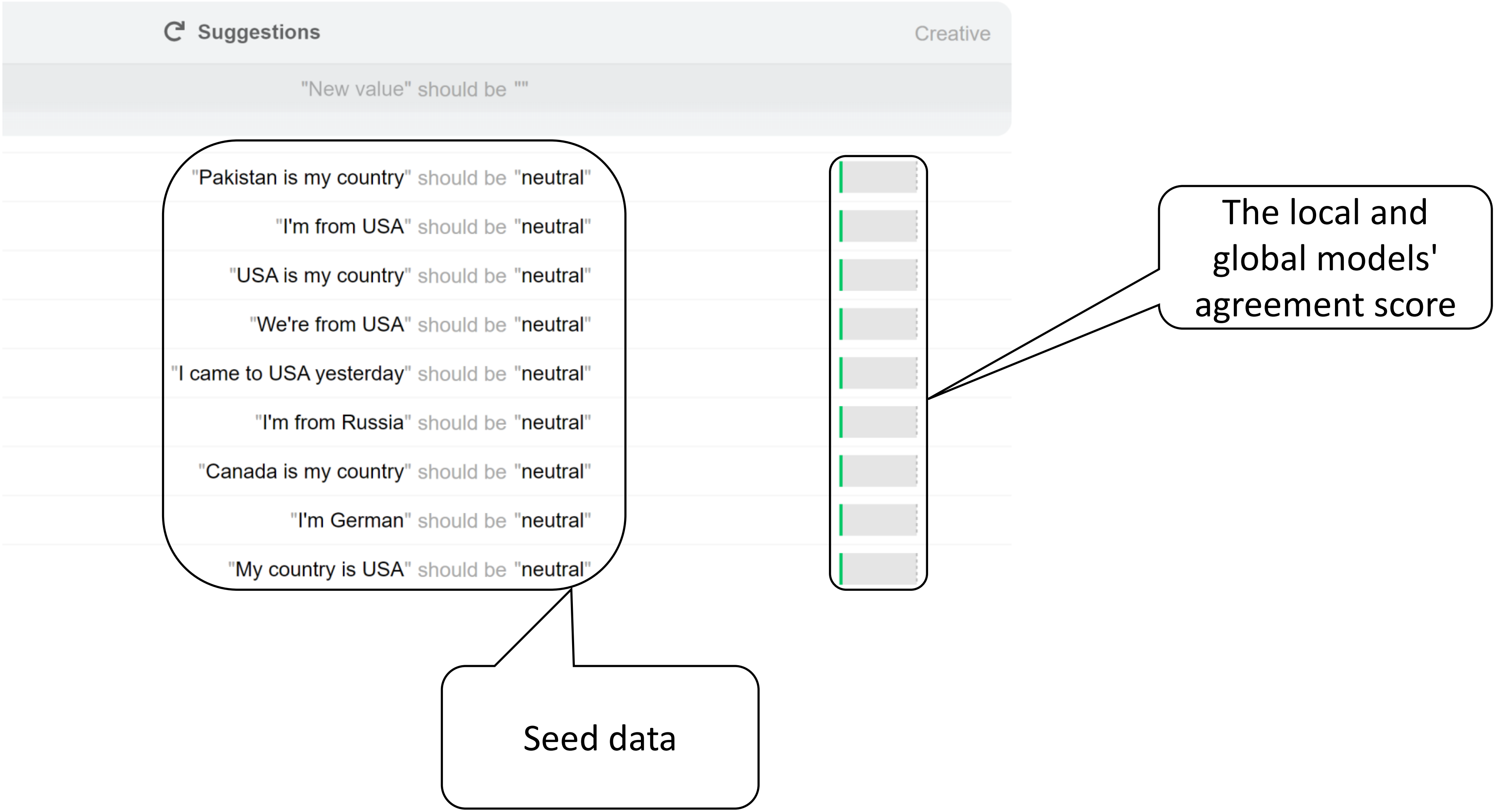
Model	MNLI											
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Avg Cluster	devtest
RoBERTa-Large	51.85	**53.57**	53.85	54.84	55.56	58.82	65.71	66.56	**68.75**	76.19	60.57	93.46
Reweighing	51.85	**53.57**	30.77	58.06	55.56	58.82	68.57	65.91	**68.75**	73.81	58.57	93.46
Paraphrasing	51.85	42.86	53.85	54.84	44.44	58.82	65.71	65.91	**68.75**	26.19	53.32	86.45
TDG (single)	51.85	**53.57**	61.54	**67.74**	**66.67**	**64.71**	65.71	**75.68**	66.67	76.19	**65.03**	-
TDG (all)	**59.26**	**53.57**	**64.28**	61.29	55.56	**64.71**	**74.28**	68.18	**68.75**	**78.57**	64.85	**93.62**

NLP demo:

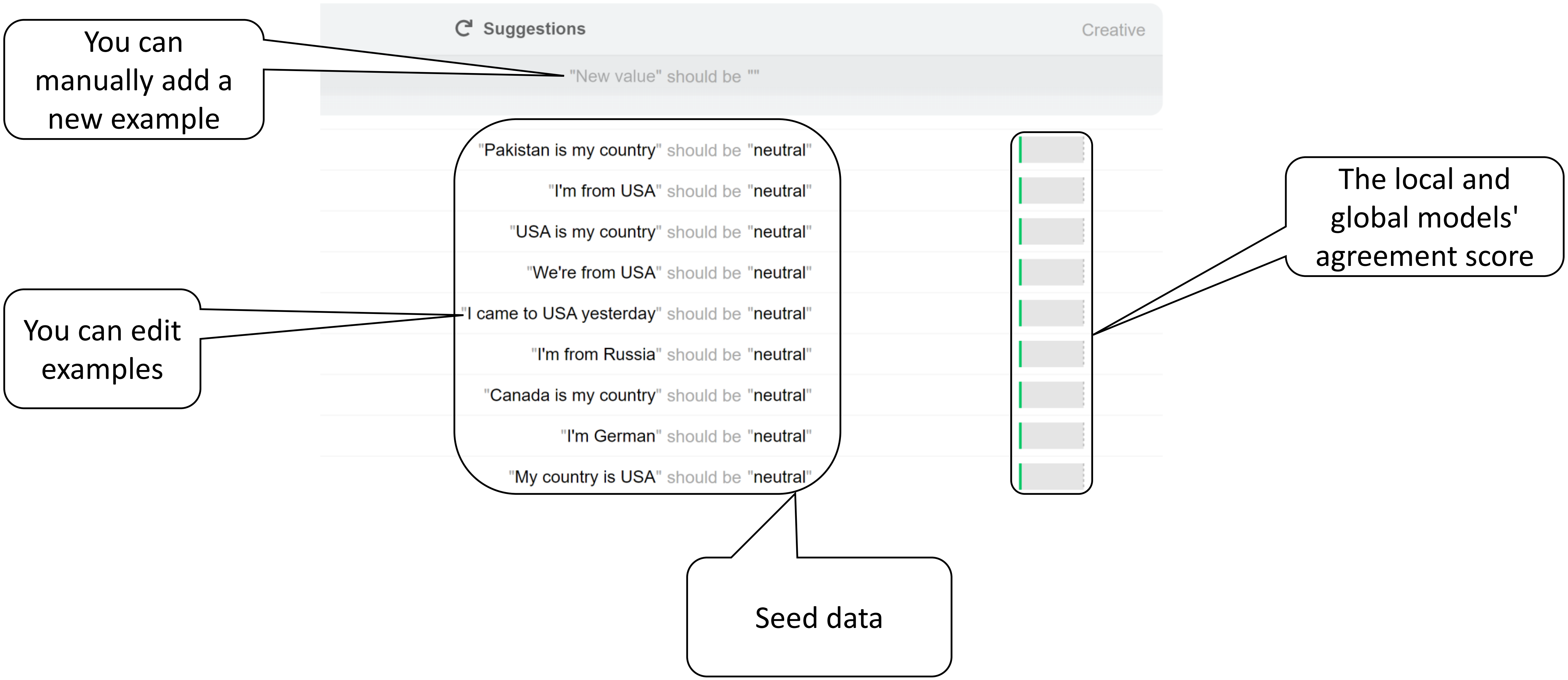
- **Goal**: checking if nationality is neutral
- **Model**: RoBerta[1] on SST[2]
- **Tool**: CoDev backend using Adatest[3] GUI

[1] Roberta: A robustly optimized bert pretraining approach. Yinhan, et al. (2019).
[2] Stanford Sentiment Treebank
[3] Adaptive Testing and Debugging of NLP Models. Ribeiro et al. (2022)

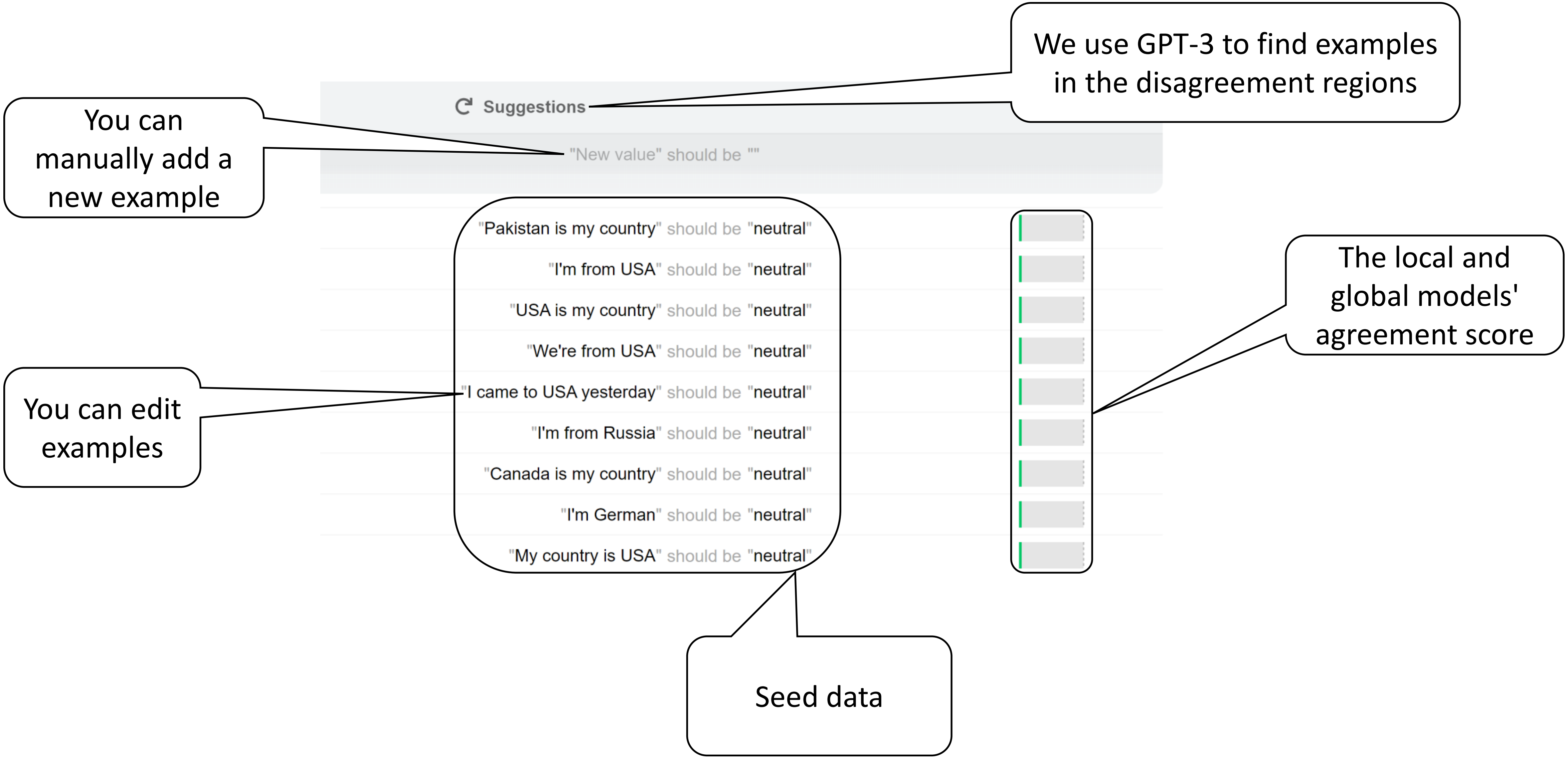
NLP demo: start from seed data



NLP demo: start from seed data

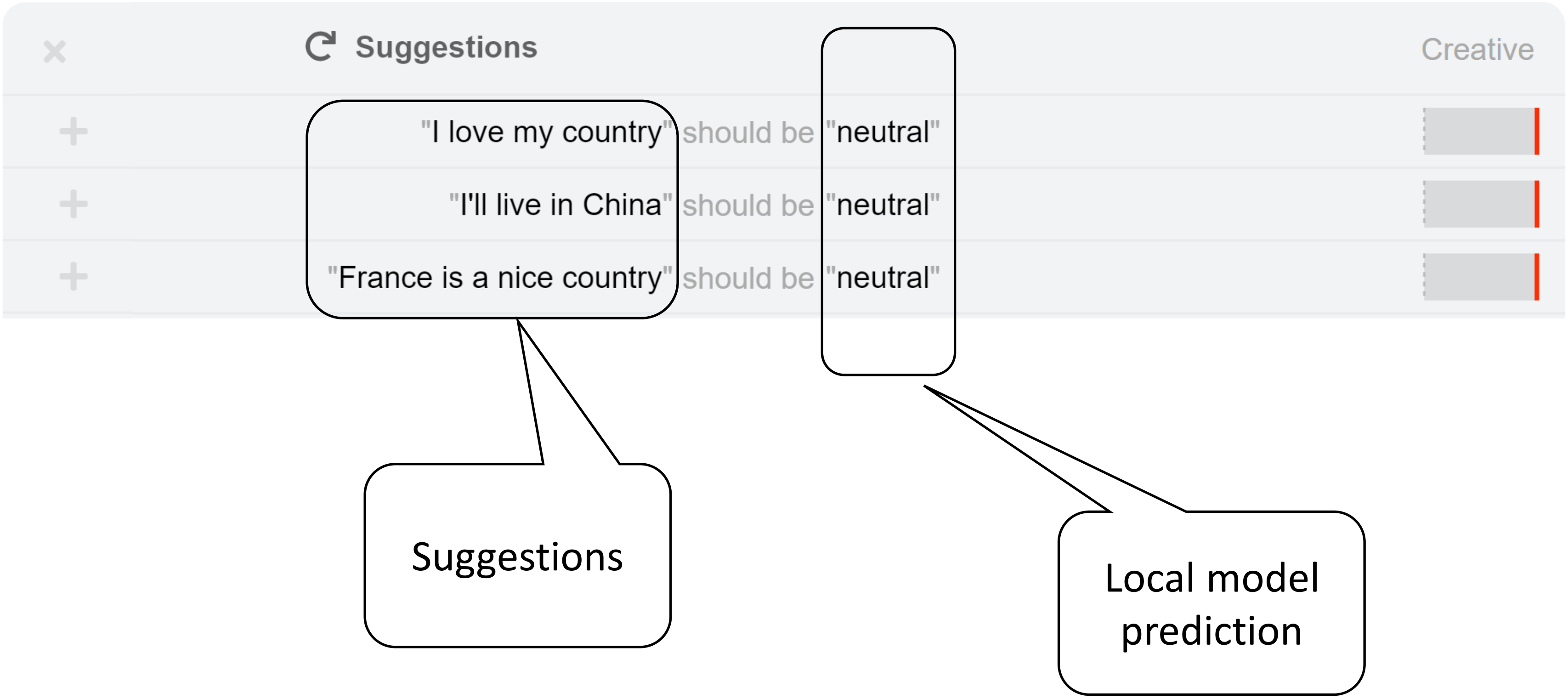


NLP demo: start from seed data

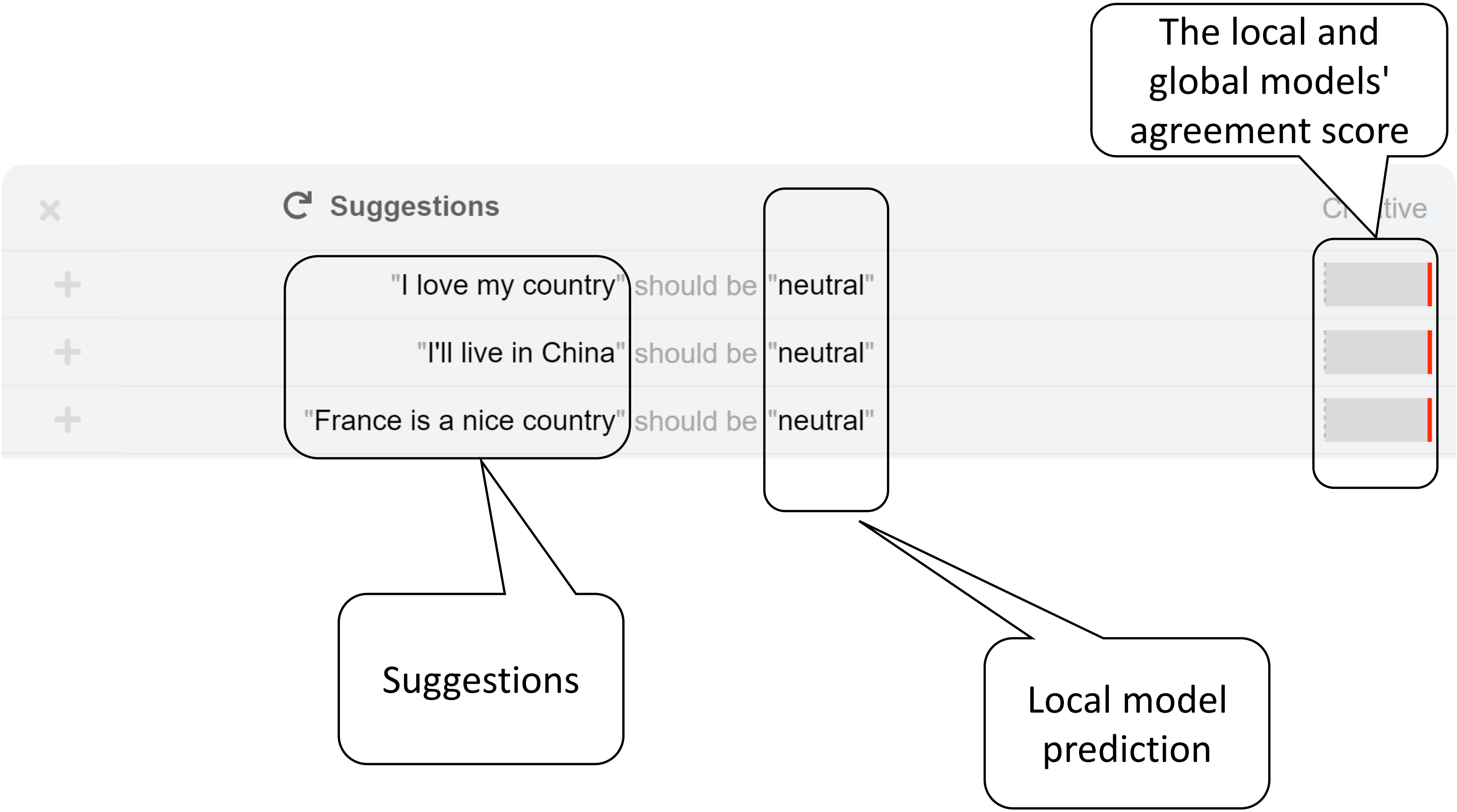


NLP demo: suggestions button generates examples on the disagreement section

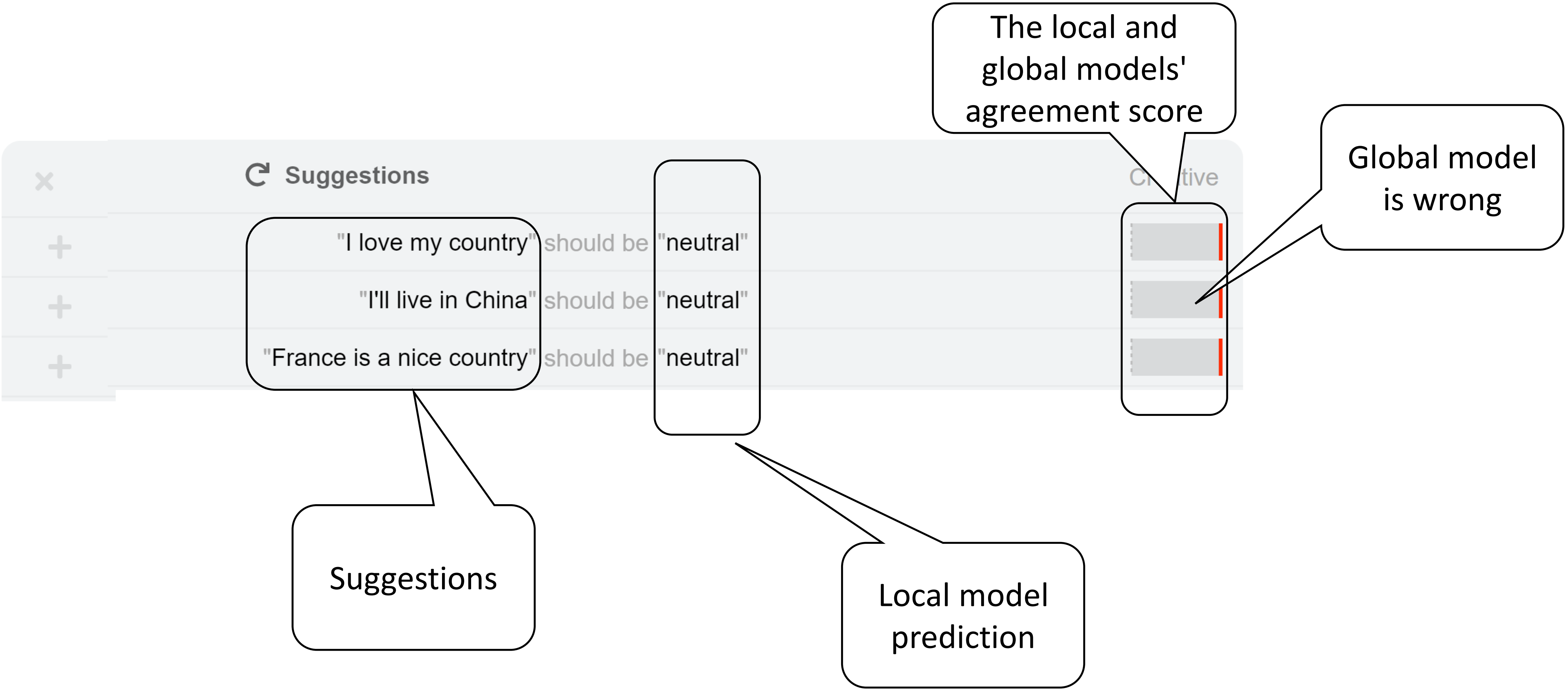
NLP demo: suggestions button generates examples on the disagreement section



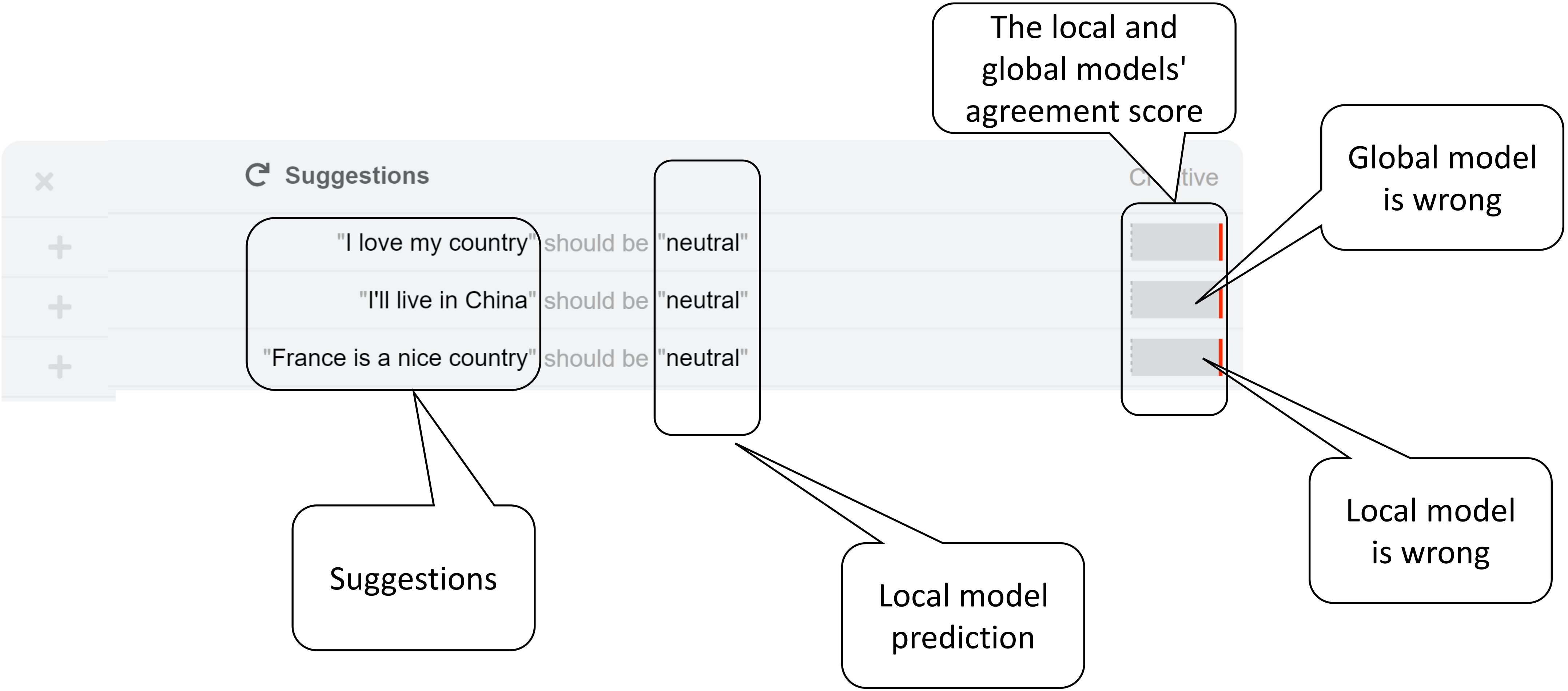
NLP demo: suggestions button generates examples on the disagreement section



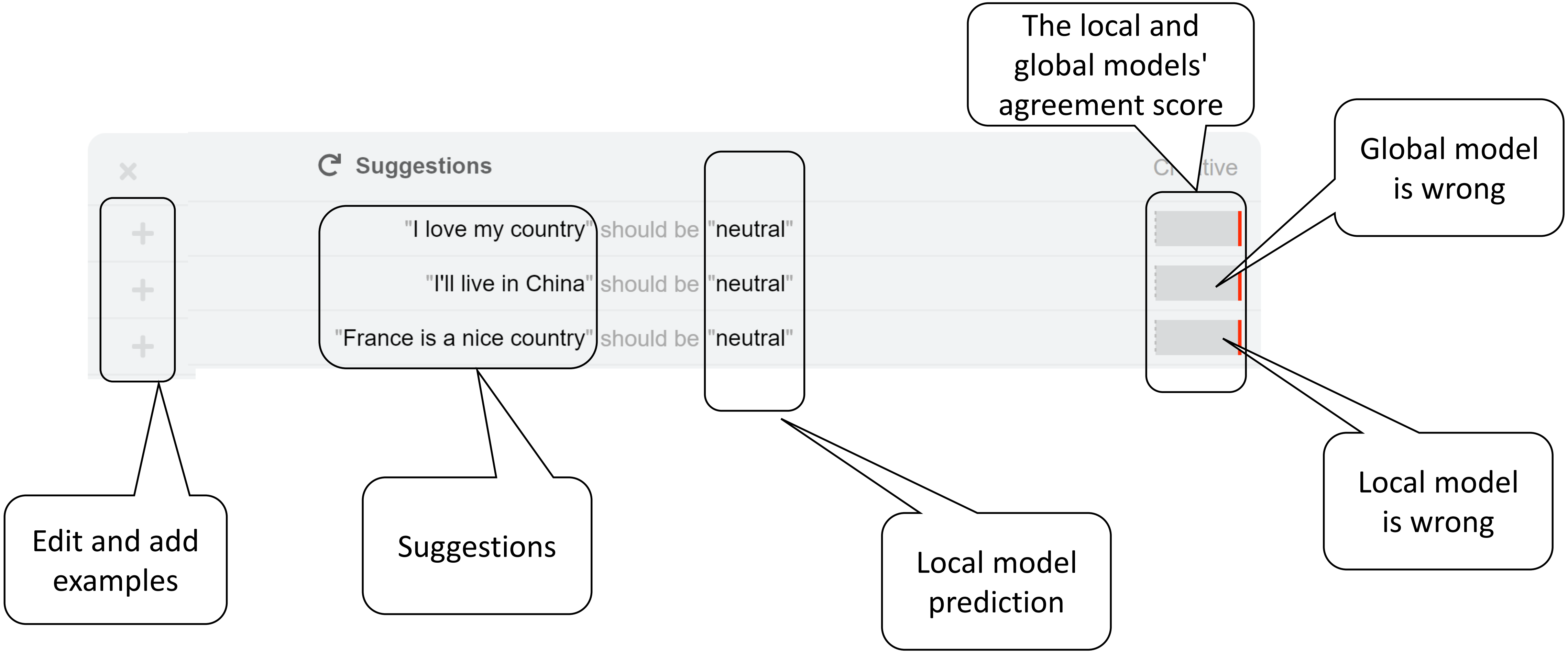
NLP demo: suggestions button generates examples on the disagreement section



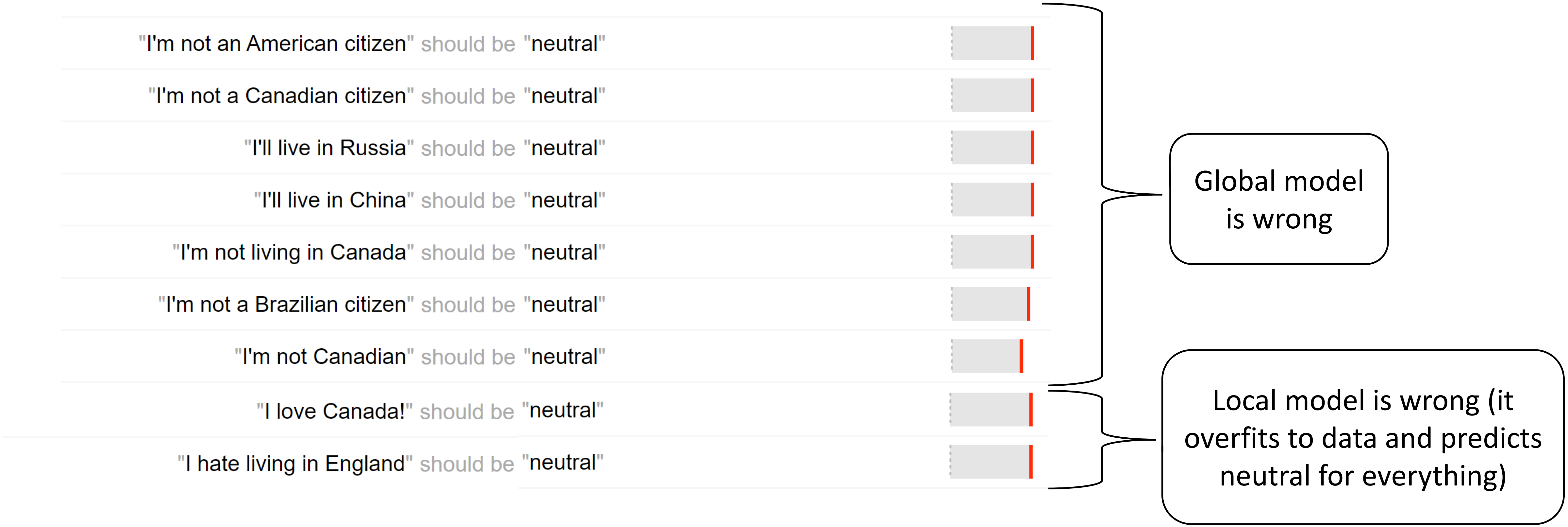
NLP demo: suggestions button generates examples on the disagreement section



NLP demo: suggestions button generates examples on the disagreement section

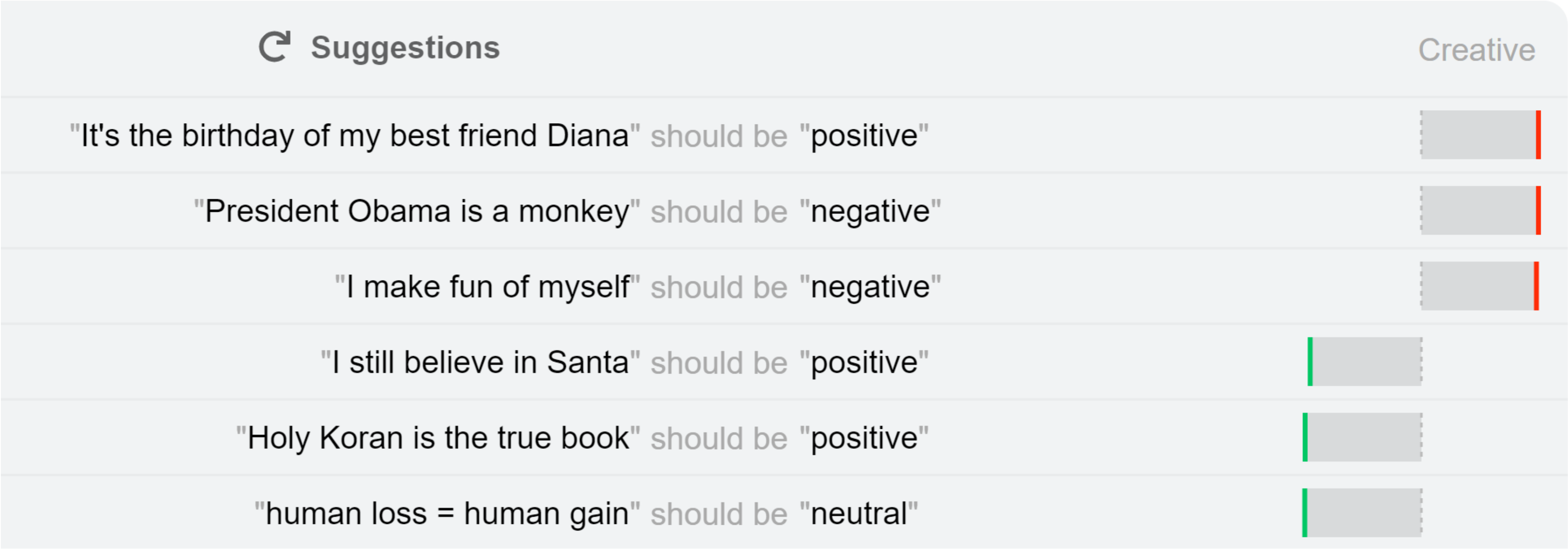


NLP demo: User keeps editing and adding new examples



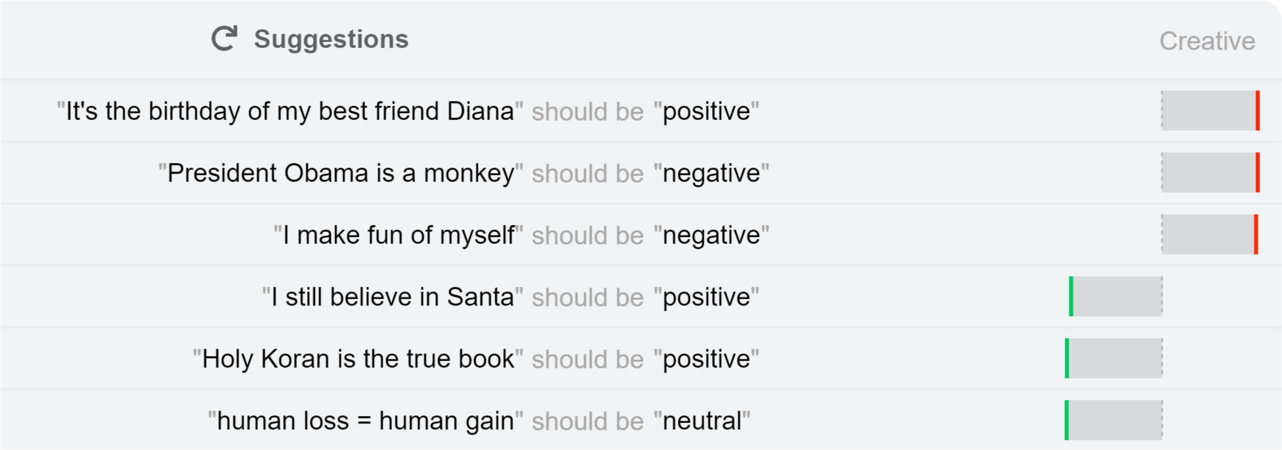
Keep Updating both models multiple times till convergence

NLP demo: Disagreements after convergence are out of domain



NLP demo: comparison with AdaTest

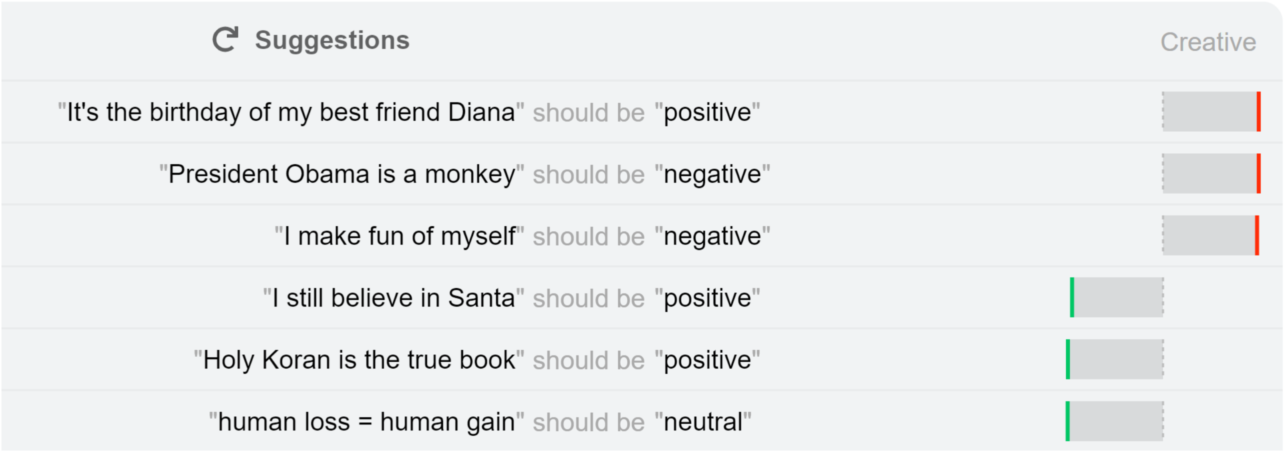
CoDev



AdaTest

NLP demo: comparison with AdaTest

CoDev

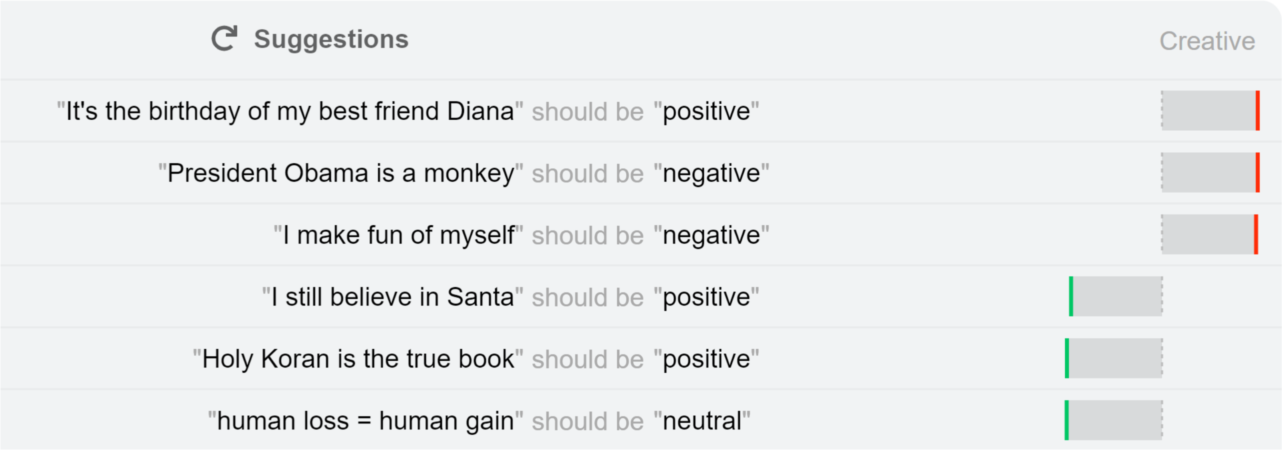


AdaTest

- Labels are predicted by local function
- Labels are less noisy and get updated as user add data
- CoDev explores buggy regions

NLP demo: comparison with AdaTest

CoDev



- Labels are predicted by local function
- Labels are less noisy and get updated as user add data
- CoDev explores buggy regions

AdaTest

- Labels are predicted by GPT3 + fraction of data
- Labels are noisy and do not get updated as user add data
- AdaTest explores correct regions instead of buggy regions

Concept	Examples	Example of bugs found by CoDev	
X person = not X person	How can I become a positive person? How can I become a person who is not negative?	predicts duplicate underfit bugs	How can I become a mysterious person? How can I become someone with no mystery?
		predicts non-duplicate overfit bugs	How can I become a blind person? How can I become someone who has lost his (physical) vision?
Modifiers changes question intent	Is Mark Wright a photographer? Is Mark Wright an accredited photographer?	predicts not-duplicate underfit bugs	Is he an artist? Is he an artist among other people?
		predicts duplicate overfit bugs	Is Joe Bennett a famous court case? Is Joe Bennett a famous American court case?

	C_{orig}: "X = not antonym (X)", C_{new}: "Modifiers changes question intent"		C_{orig}: "X = synonym (X)", C_{new}: "less X = more antonym (X)"	
	CoDev	AdaTest	CoDev	AdaTest
broken by new concept	7/50	24/50	9/50	18/50
fixed by new concept	5/50	2/50	20/50	18/50