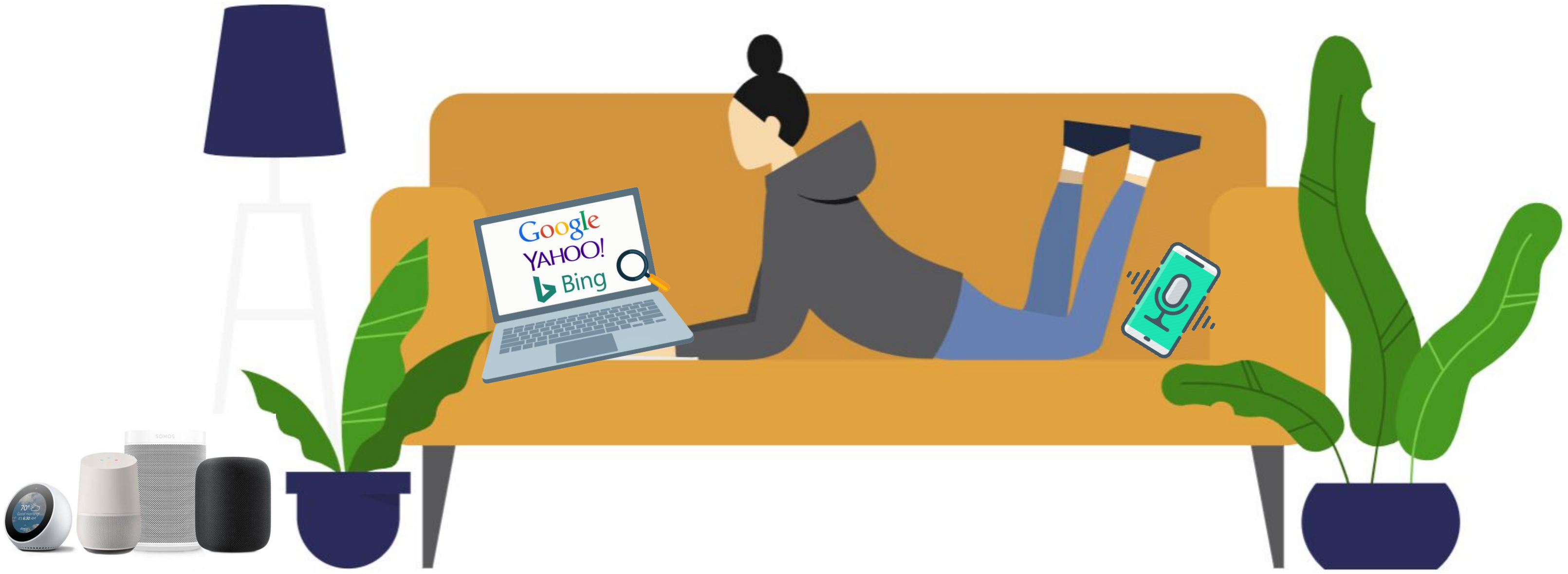
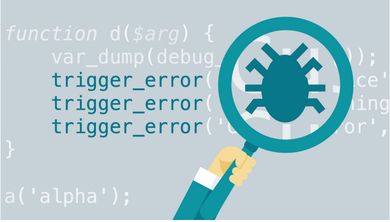
Explanation Selection Through The Lens of Free-Text and Contrastive Explanations

Ana Marasović

Allen Institute for AI (AI2) × AllenNLP × University of Washington

Natural Language Processing has become an integral part of most people's daily lives





Technically robust and safe

Allows acknowledging and evaluating trade-offs

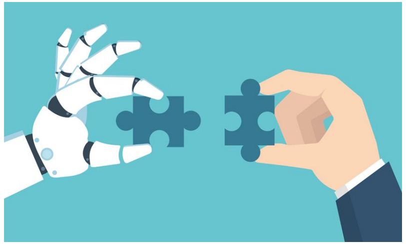
Respects quality and integrity of data

Trustworthy AI Contracts

Encourages green AI

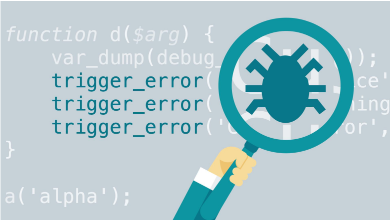
Supports users’ agency and oversight

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI

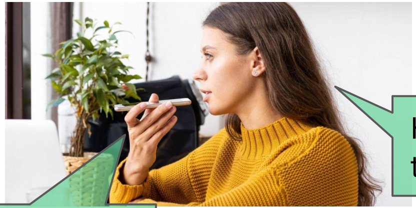
Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.



Technically robust and safe

Allows acknowledging and evaluating trade-offs

Why this answer?



How to change the answer?

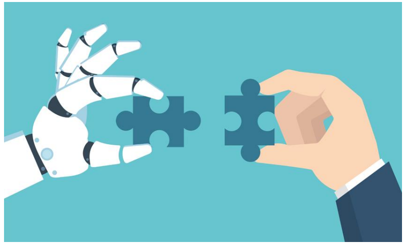
What if I change the input in this way?

Respects quality and integrity of data

Encourages green AI

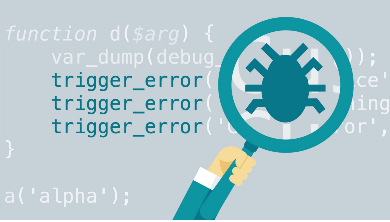
Supports users' agency and oversight

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI

Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.

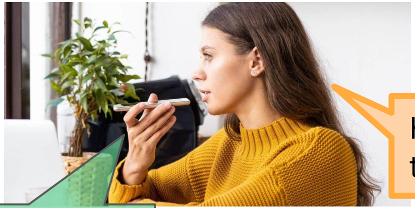


Technically robust and safe

Why this answer?

Allows acknowledging and evaluating trade-offs

Respects quality and integrity of data



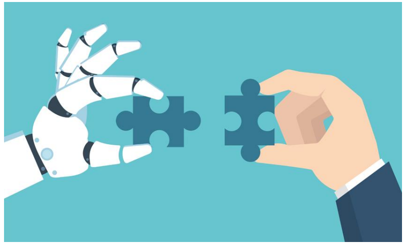
How to change the answer?

Encourages green AI

What if I change the input in this way?

Supports users' agency and oversight

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI

Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.

One approach to realizing some of the trustworthy AI goals is via **local explanations**: justifications of models' individual predictions

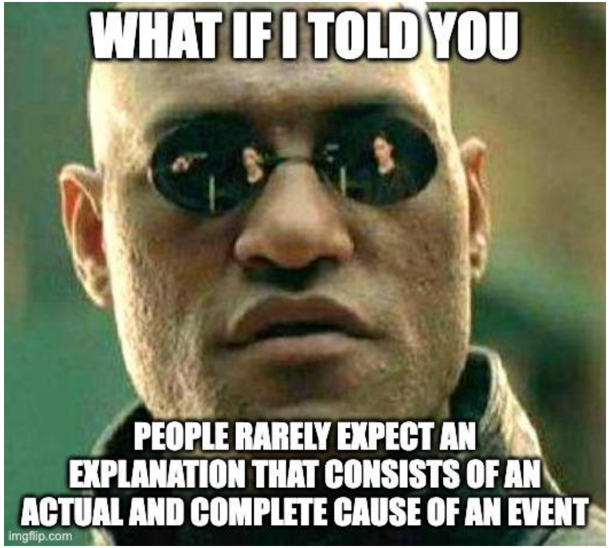
A dominant ML/NLP perspective on local explanations

→ Causal attribution: given a set of factors (usually, input tokens/pixels), select ***all factors*** that ***cause*** the model's decision

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

A dominant ML/NLP perspective on local explanations

→ Causal attribution: given a set of factors (usually, input tokens/pixels), select ***all factors*** that ***cause*** the model's decision



Miller’s 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

1. **Cognitive load**: causal chains are often too large to comprehend

2. Explainee cares only about a small number of causes (relevant to the context)

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

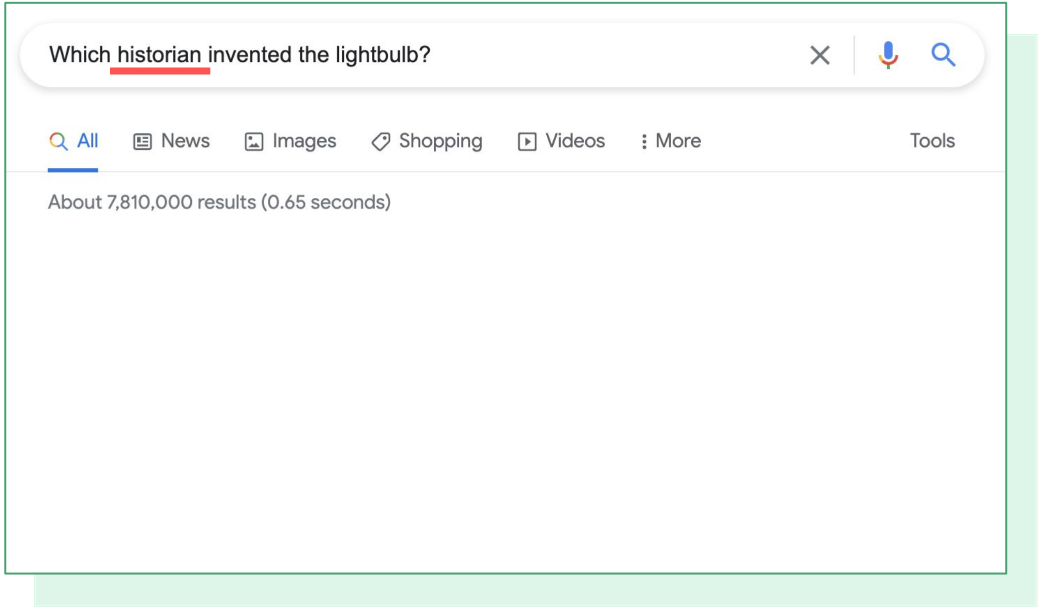
Miller's 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

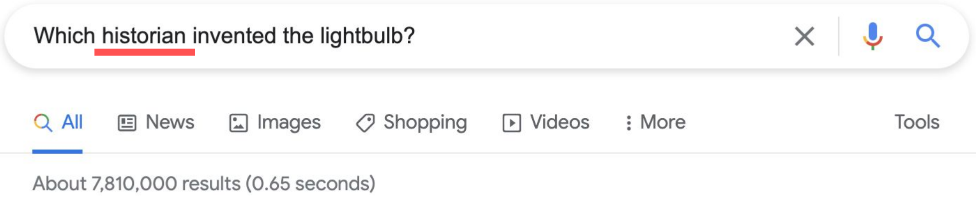
1. **Cognitive load**: causal chains are often too large to comprehend

2. Explainee cares only about a small number of causes (relevant to the context)

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.



Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.



None *because* Thomas Edison is credited as the primary inventor of the lightbulb and Edison was not a historian

Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.

Thomas Alva Edison (February 11, 1847 – October 18, 1931) was an American inventor and businessman who has been described as America's greatest inventor.[1][2][3] He developed many devices in fields such as electric power generation, mass communication, sound recording, and motion pictures.[4] These inventions, which include the phonograph, the motion picture camera, and early versions of the electric light bulb, have

Thomas Edison is credited as the primary inventor of the lightbulb and Edison was not a historian.

Answering “why” by highlighting

Sylvester Stallone has made some crap films in his lifetime, but this has got to be one of the worst. A totally dull story that thinks it can use various explosions to make it interesting, "the specialist" is about as exciting as an episode of "dragnet," and about as well acted. Even some attempts at film noir mood are destroyed by a sappy script, stupid and unlikable characters, and just plain nothingness. Who knew a big explosion could be so boring and anti-climactic?

`Label:` negative sentiment

Introduced in Zaidan et al. Using “Annotator Rationales” to Improve Machine Learning for Text Categorization. NAACL 2007.
Popularized with Lei et al. Rationalizing Neural Predictions. EMNLP 2016.

Answering "why" by highlighting

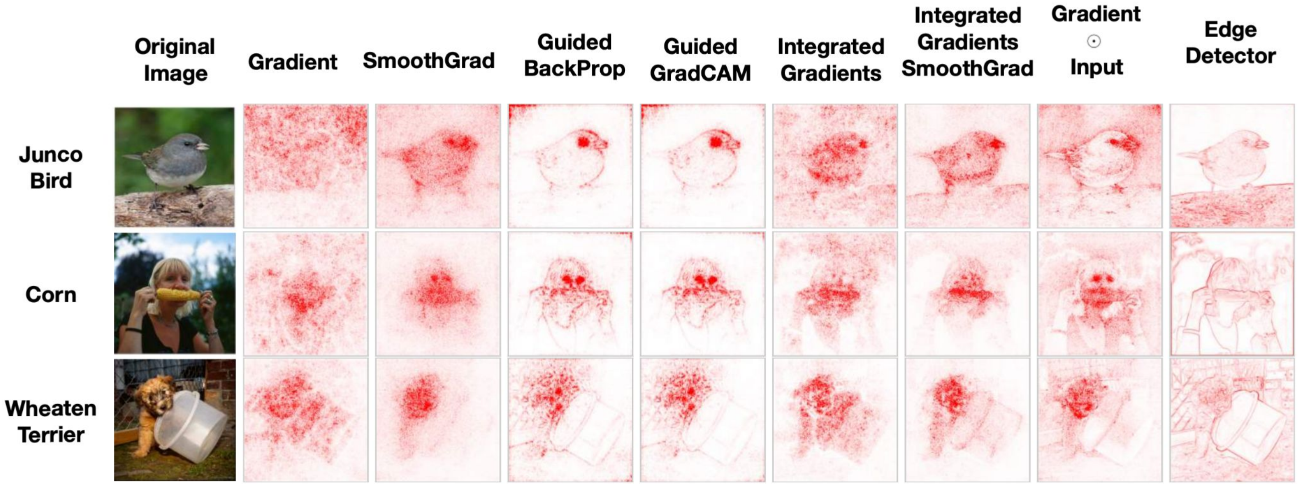
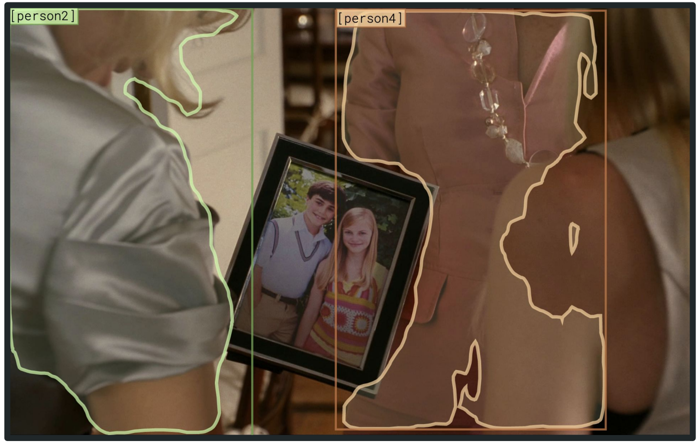


Figure from Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018.

Cognitive load of understanding highlighting is very high when the reason is not explicitly stated in the input



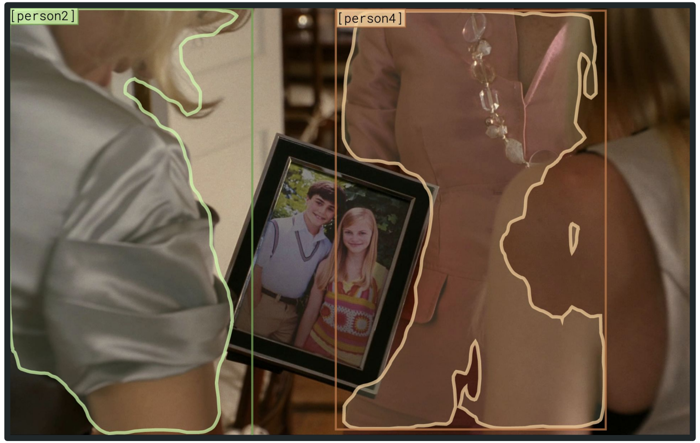
Question: What is going to happen next?

Answer: [person2] holding the photo will tell [person4] how cute their children are.

Free-text explanation: It looks like [person4] is showing the photo to [person2], and they will want to be polite.

Example from From Zellers et al. Recognition to Cognition: Visual Commonsense Reasoning. CVPR 2019.

Cognitive load of understanding highlighting is very high when the reason is not explicitly stated in the input



Free-text explanation:

- [person4] is showing the photo to [person2]
- [person2] will want to be polite

We cannot highlight this in the input!

Example from From Zellers et al. Recognition to Cognition: Visual Commonsense Reasoning. CVPR 2019.

Miller's 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

1. **Cognitive load**: causal chains are often too large to comprehend

2. Explainee cares only about a small number of causes (relevant to the context)

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

Miller's 2nd Insight from Social Science

Explanations are **contrastive** = responses to:

"Why P rather than Q?"

"How to change the answer from P to Q?"

where **P** is an observed event **(fact)**, and **Q** an imagined, counterfactual event that did not occur **(foil)**

Why did she get 20x less limit?

1. Make joint tax returns
2. Live in a community-property state
3. Be married for a long time
4. ....

What are the factors in the application that would need to change to get the same limit?
woman → ?

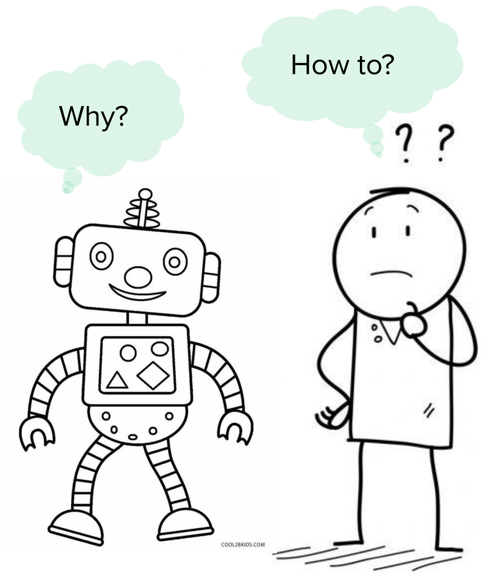
Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

“Understanding how people define, generate, select, evaluate, and present explanations seems almost essential”

People assign human-like traits to AI models (**anthropomorphic bias**)

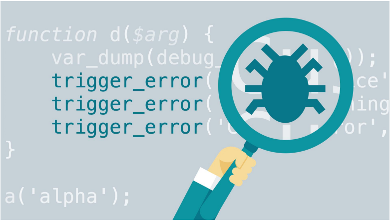
⇒ People expect explanations of models’ behavior to follow the same conceptual framework used to explain human behavior

⇒ No users’ agency otherwise



Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

Trustworthy AI Contracts



Technically robust and safe

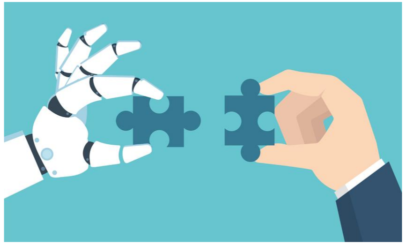
Respects quality and integrity of data

Allows acknowledging and evaluating trade-offs

Supports users' agency and oversight

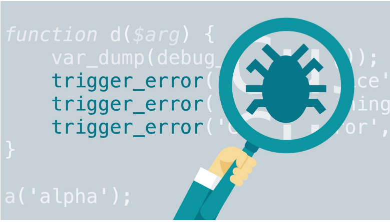
Encourages green AI

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI

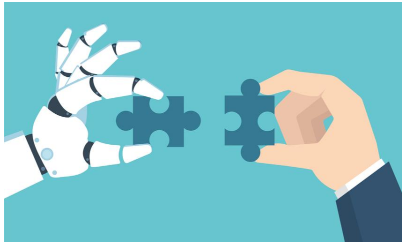
Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.



Technically robust and safe

Respects quality and integrity of data

Supports users' agency and oversight



Allows assessing the impact on individuals, society, democracy

Encourages green AI

Allows acknowledging and evaluating trade-offs

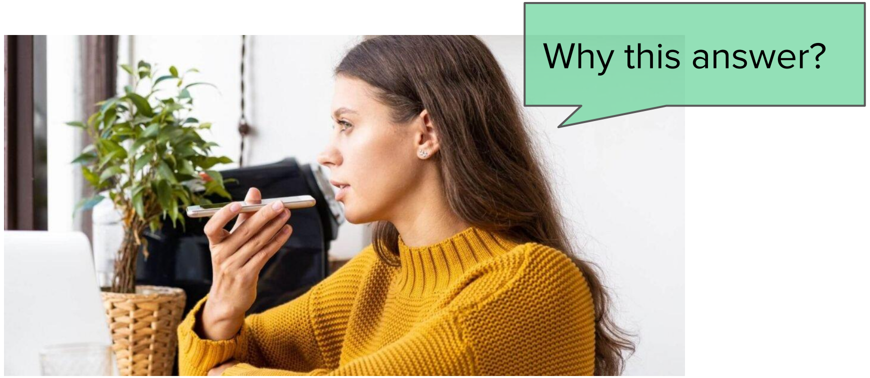
Why this answer?

How to change the answer?

What if I change the input in this way?

European ethics guidelines for trustworthy AI

Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.



Data

Wiegreffe* and **Marasović* (equal contributions)**. Teach Me to Explain: A Review of Datasets for Explainable NLP. NeurIPS 2021.

Modeling

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

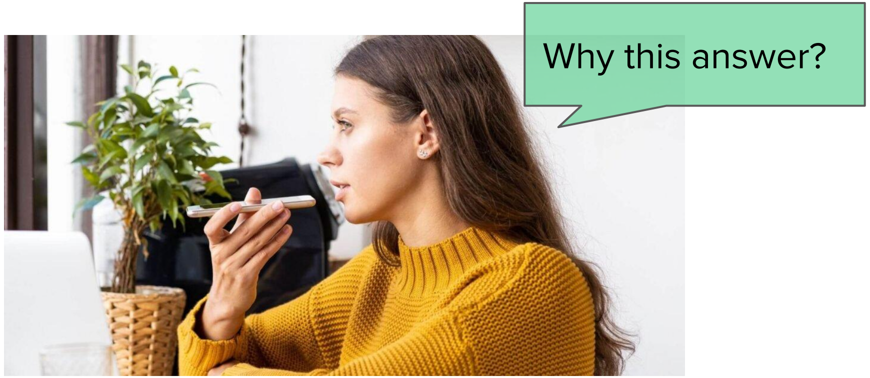
Marasović*, Beltagy*, et al. Few-Shot Self-Rationalization with Natural Language Prompts. arXiv 2021.

Theoretical and Empirical Evaluation

Wiegreffe, **Marasović**, Smith. Measuring Association Between Labels and Free-Text Rationales. EMNLP 2021.

Sun and **Marasović**. Effective Attention Sheds Light On Interpretability. Findings of ACL 2021.

Jacovi, **Marasović**, et al. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. FAccT 2021.



Data

Wiegreffe* and **Marasović* (equal contributions)**. Teach Me to Explain: A Review of Datasets for Explainable NLP. NeurIPS 2021.

Modeling

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Marasović*, Beltagy*, et al. Few-Shot Self-Rationalization with Natural Language Prompts. arXiv 2021.

Theoretical and Empirical Evaluation

Wiegreffe, **Marasović**, Smith. Measuring Association Between Labels and Free-Text Rationales. EMNLP 2021.

Sun and **Marasović**. Effective Attention Sheds Light On Interpretability. Findings of ACL 2021.

Jacovi, **Marasović**, et al. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. FAccT 2021.

Explaining Visual Reasoning

Marasović et al (2020)

Natural Language Rationales with Full-Stack Visual Reasoning:
From Pixels to Semantic Frames to Commonsense Graphs

Explaining reasoning requires more than highlighting

Question: Where is a frisbee in play likely to be?

Answer choices: outside, park, roof, tree, <u>air</u>

Free-text explanation: A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game so while in play it is most likely to be in the air.

Aggarwal et al. (2021)

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

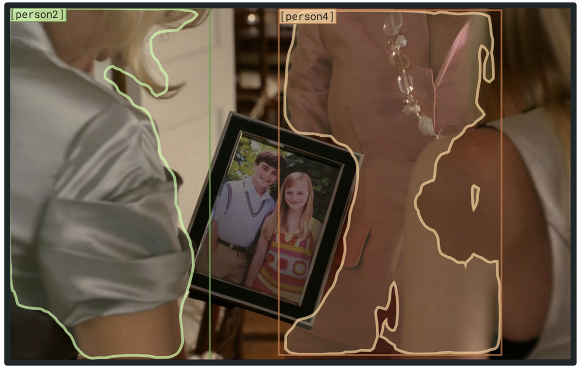
Explaining reasoning requires more than highlighting

Question: Where is a frisbee in play likely to be?

Answer choices: outside, park, roof, tree, <u>air</u>

Free-text explanation: A frisbee is a concave plastic disc designed for skimming through the air as an outdoor game so while in play it is most likely to be in the air.

Aggarwal et al. (2021)



Question: What is going to happen next?

Answer: [person2] holding the photo will tell [person4] how cute their children are.

Free-text explanation: It looks like [person4] is showing the photo to [person2], and they will want to be polite.

Zellers et al. (2019)

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning. Findings of EMNLP 2020.

How to generate free-text explanations?

Step 1:

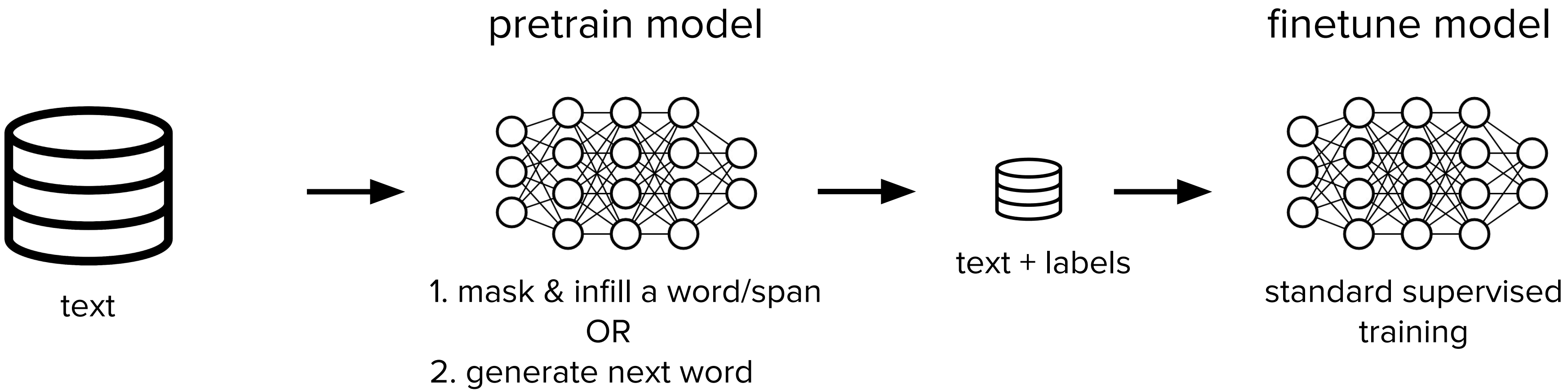
Find some human-written explanations[◊]

Step 2:

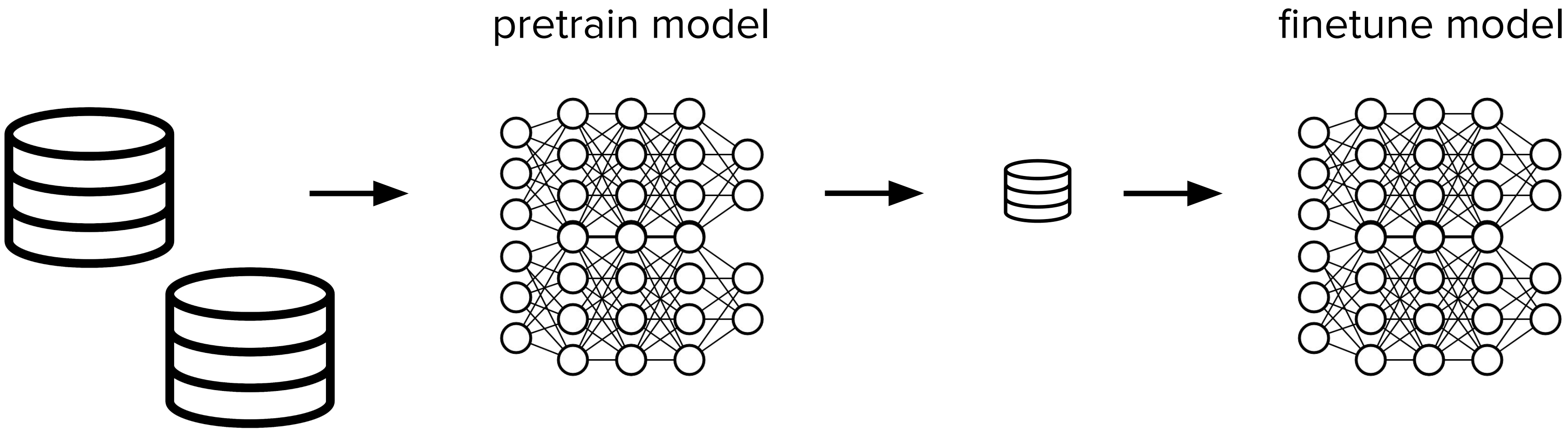
Finetune a pretrained transformer-based generation models (T5, GPT-2/Neo)

[◊] Wiegreffe* and **Marasović***. Teach Me to Explain: A Review of Datasets for Explainable NLP. NeurIPS 2021.

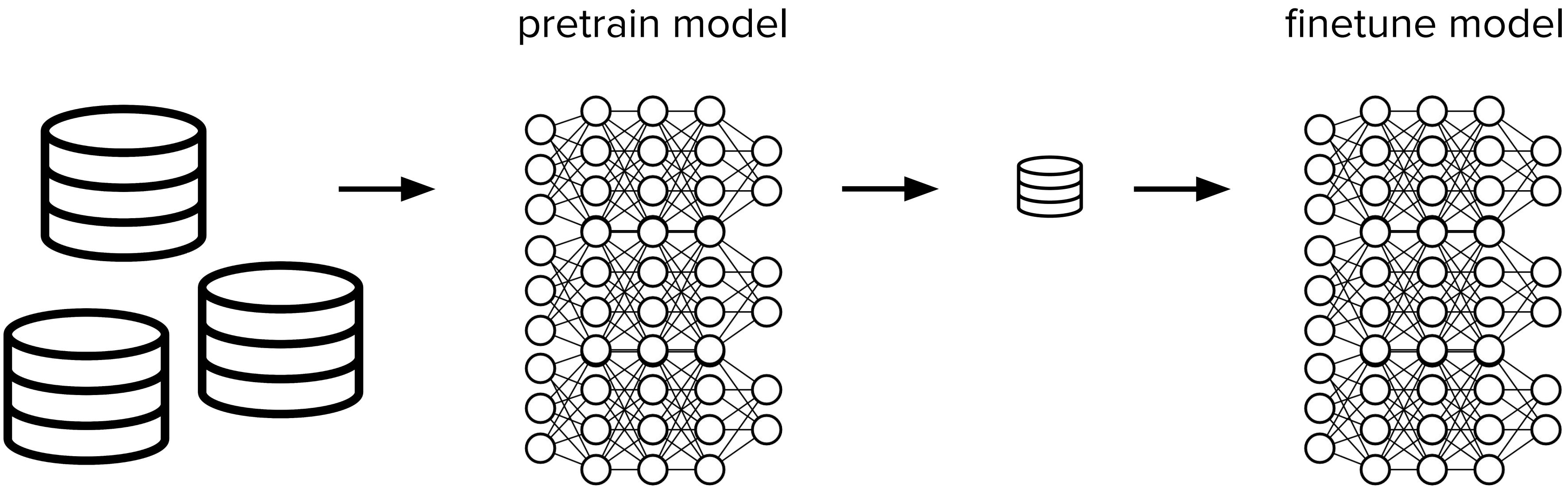
Pretrain-Finetune Paradigm



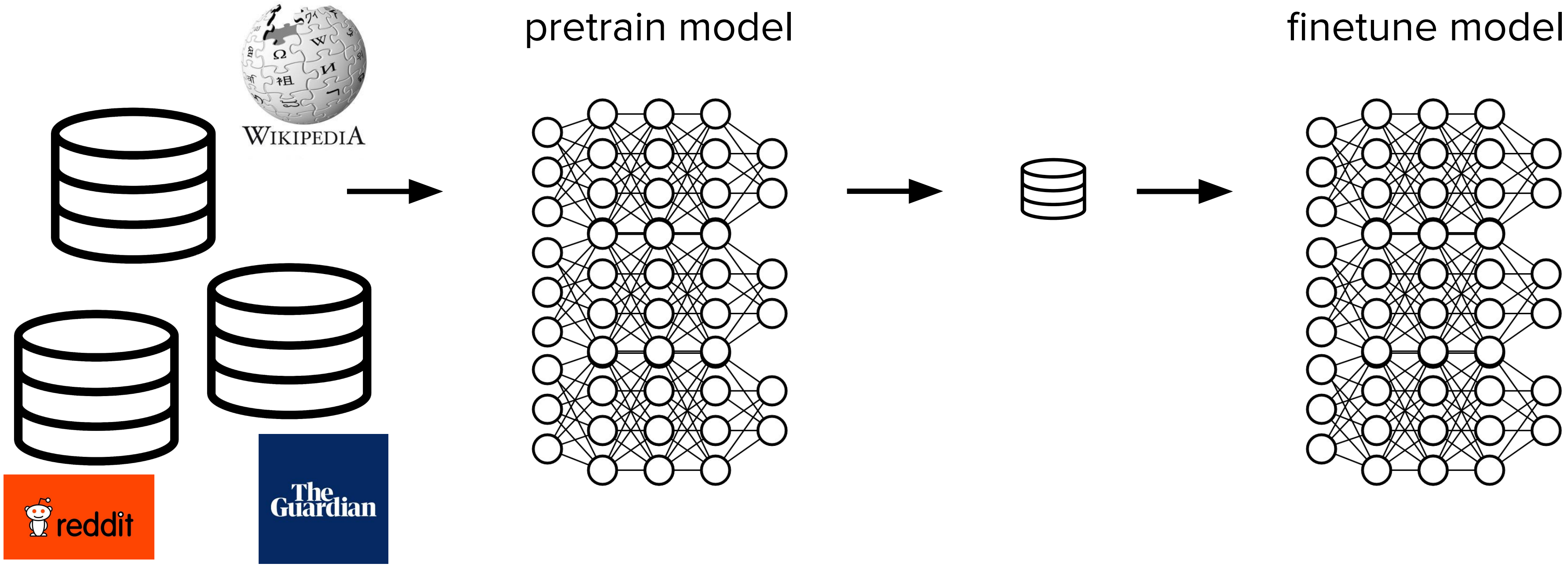
Pretrain-Finetune Paradigm



Pretrain-Finetune Paradigm



Pretrain-Finetune Paradigm



Dodge, Sap, **Marasović**, Agnew, Ilharco, Groeneveld, Mitchell, Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. EMNLP 2021.

How to generate free-text explanations?

Step 1:

Find some human-written explanations[◊]

Step 2:

Finetune a pretrained transformer-based generation models (T5, GPT-2/Neo)

[◊] Wiegreffe* and **Marasović***. Teach Me to Explain: A Review of Datasets for Explainable NLP. NeurIPS 2021.

Transformer

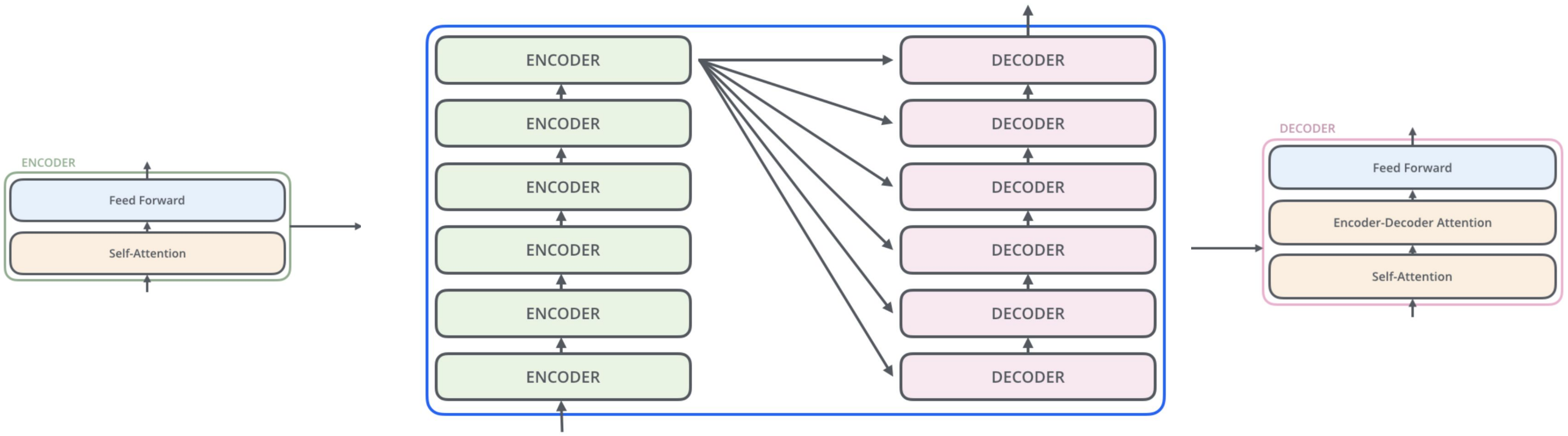
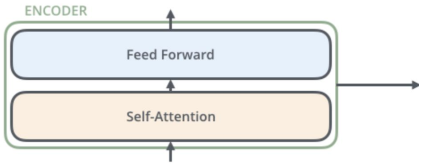
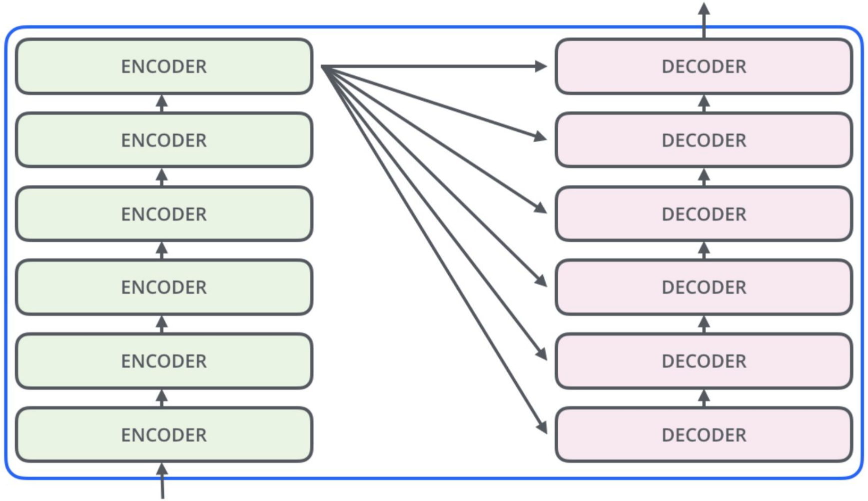
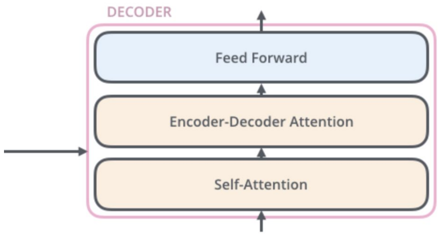


Figure from Alammar. The Illustrated Transformer.

Generating Explanations





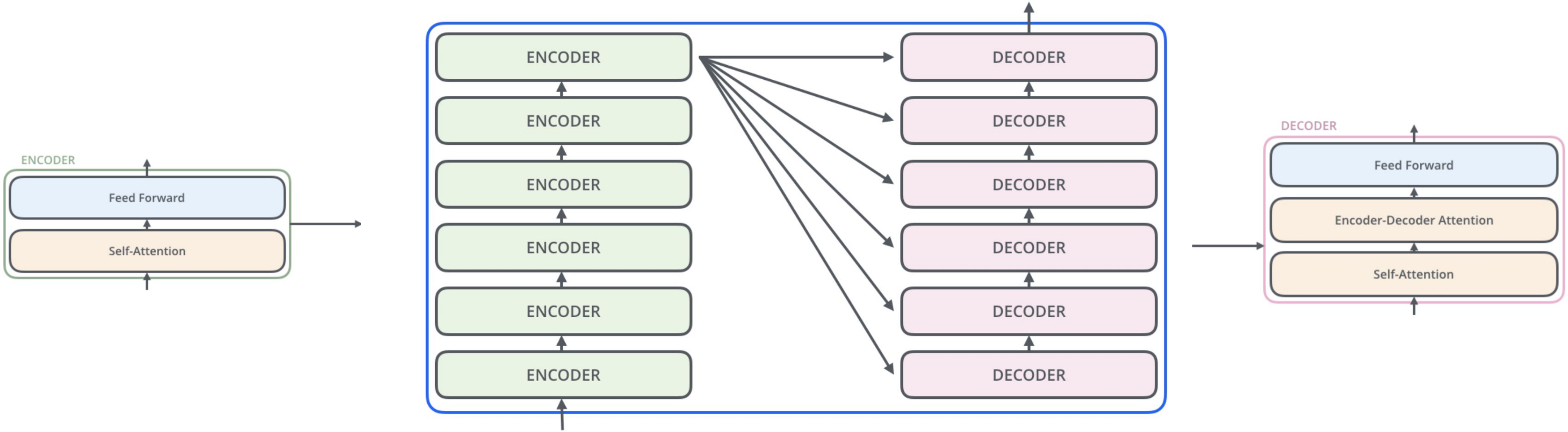



```
question: where is a frisbee in play likely
to be? choice: outside choice: park choice:
roof choice: tree choice: air
```

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

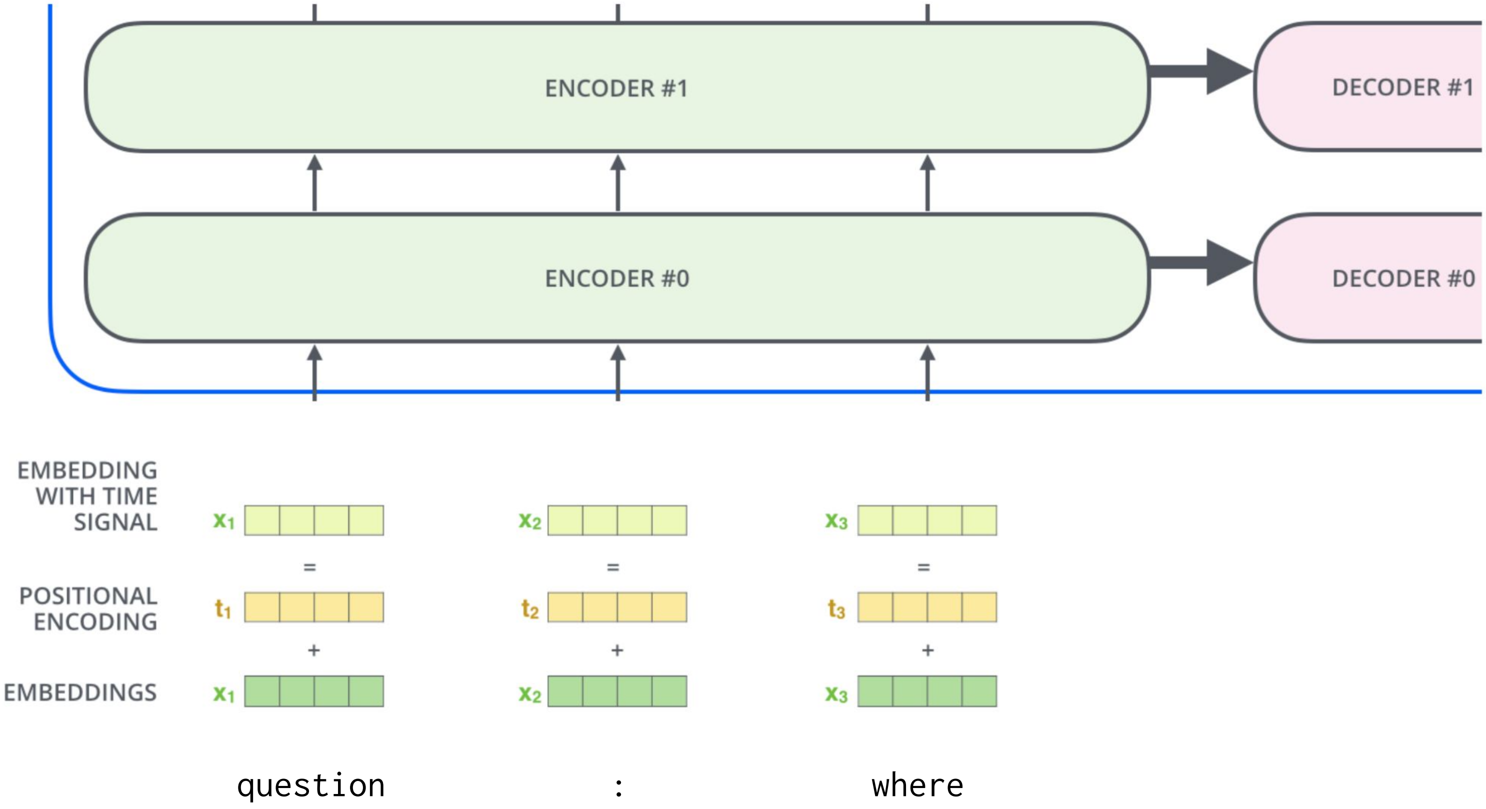
Generating Explanations

Air because a frisbee is a concave plastic disc designed for skimming through the air as an outdoor game so while in play it is most likely to be in the air.

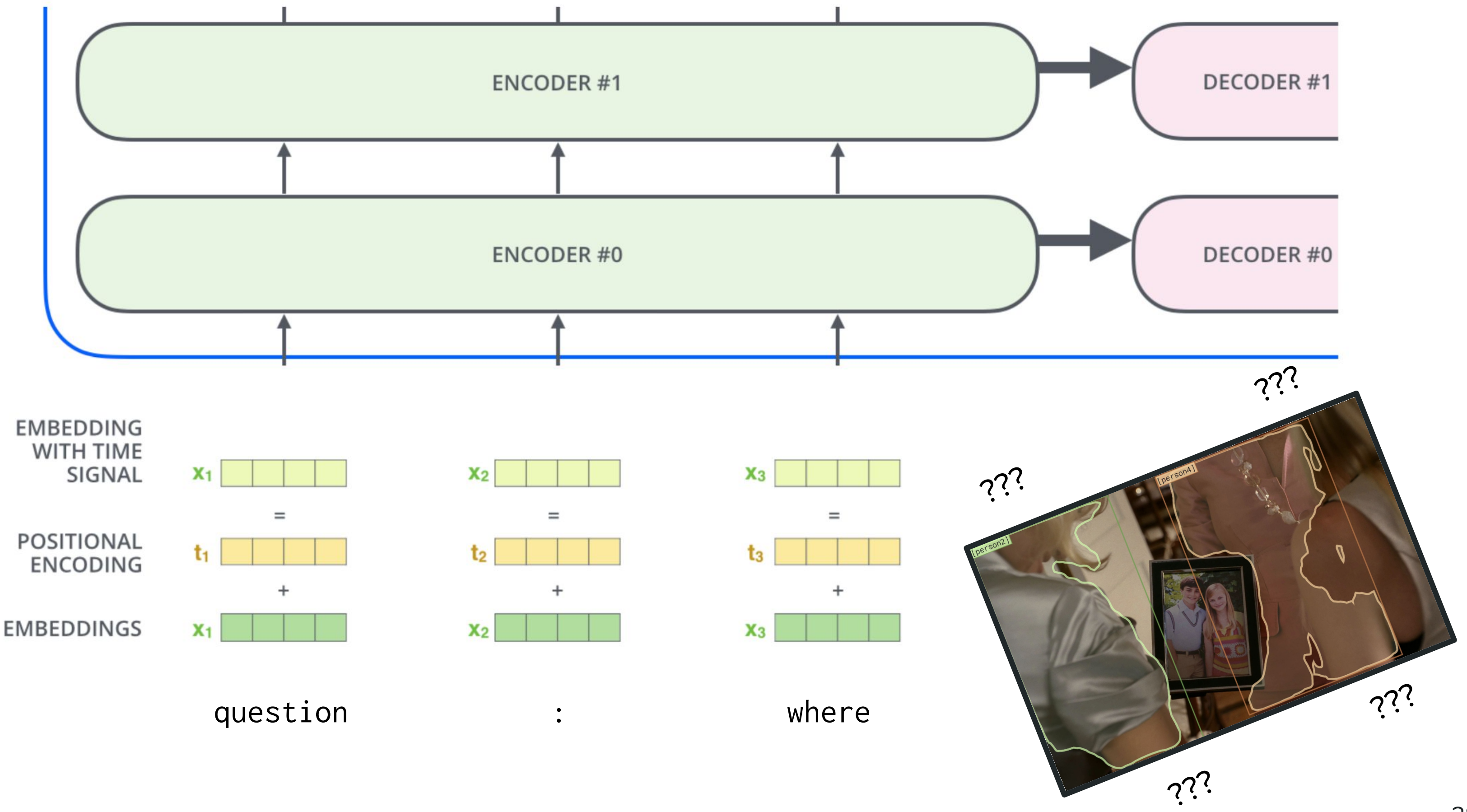


question: where is a frisbee in play likely to be? choice: outside choice: park choice: roof choice: tree choice: air

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.



Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.



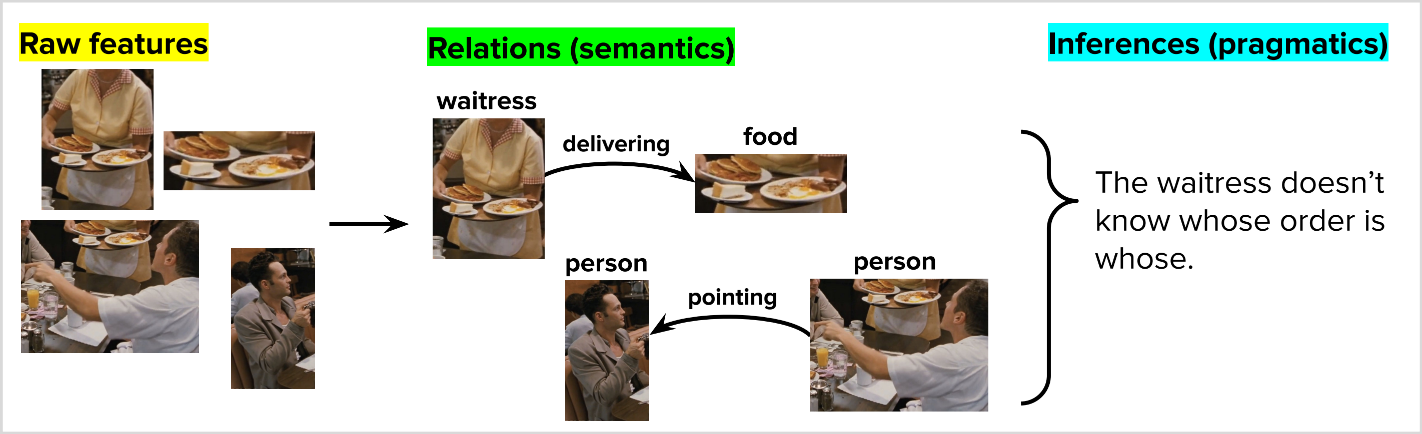
Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Key challenge: image representation beyond explicit content

Question: Why is person on the right pointing to the person on the left?

Answer: He is telling the waitress that the person on the left ordered the pancakes.

Free-text explanation: She is delivering food to the table and she doesn't know whose order is whose.



Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

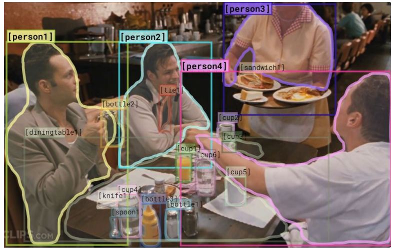
Raw features

Relations (semantics)

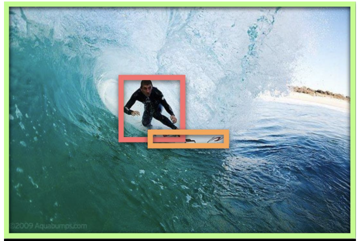
Inferences (pragmatics)

The waitress doesn't know whose order is whose.

object detection◊



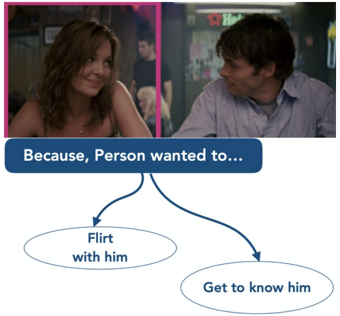
grounded situation recognition○



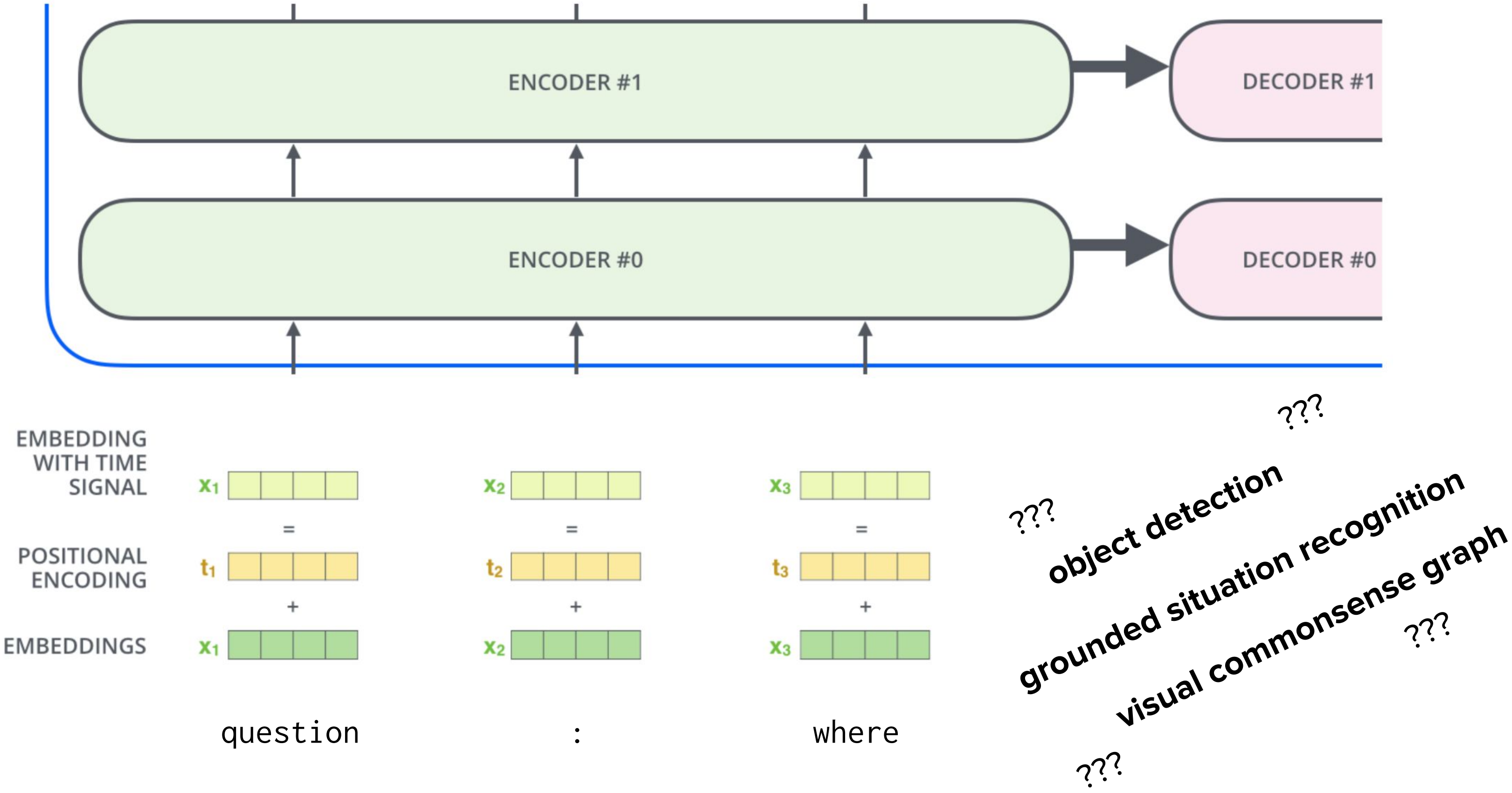
Surfing			
Agent	Tool	Path	Place
Man	Surfboard	Water	Ocean

visual commonsense graph□

"order a drink"

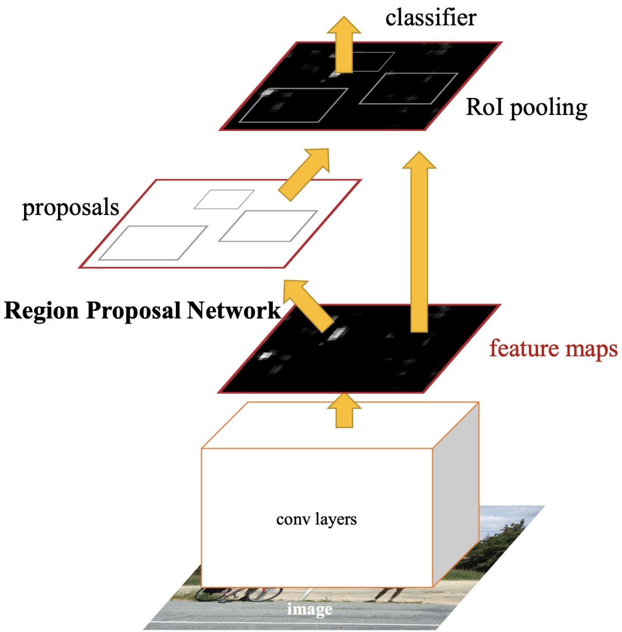


◊ Ren et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2015.
○ Pratt et al. Grounded Situation Recognition. ECCV 2020.
□ Park et al. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. ECCV 2020.



Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Back to Basics: Object Detection



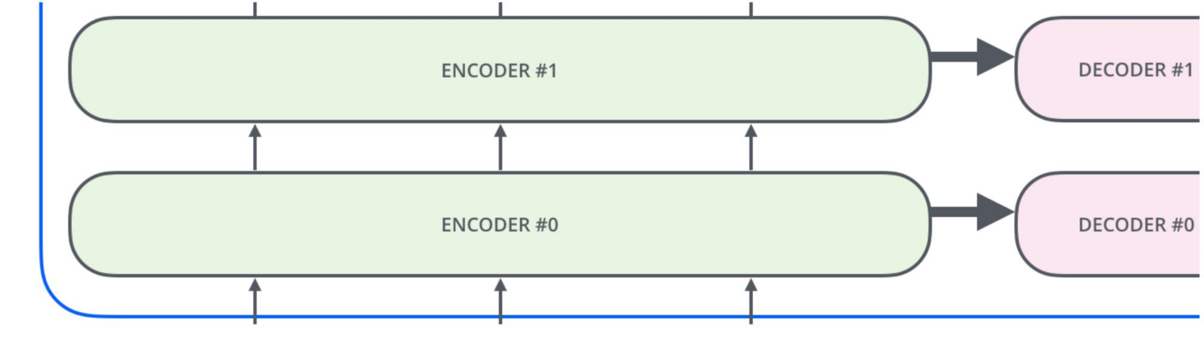
Output:

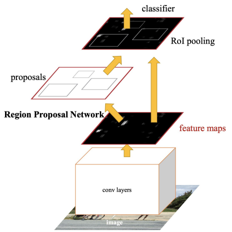
1. class (e.g. cup)
2. vector

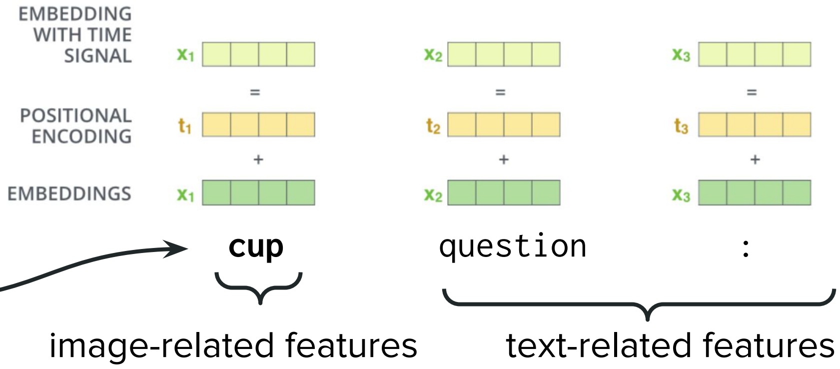
for each detected object

Figure from Ren et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2015.

Uniform fusion: Prepend **object labels** to text





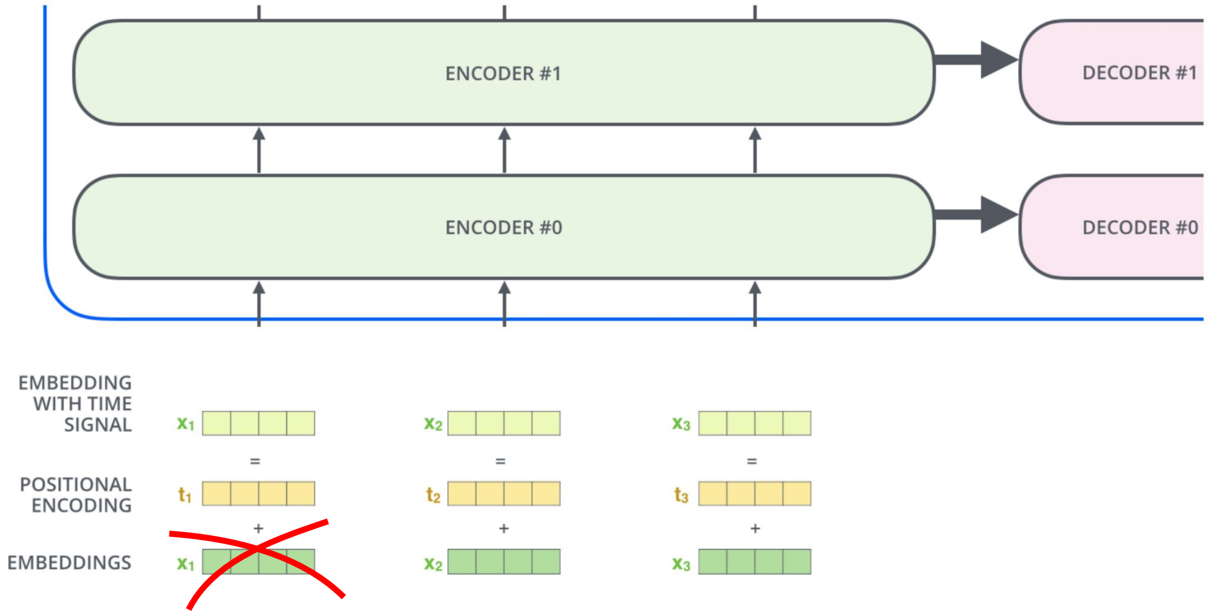


Pro: very simple

Con: prone to propagation of errors from external vision models

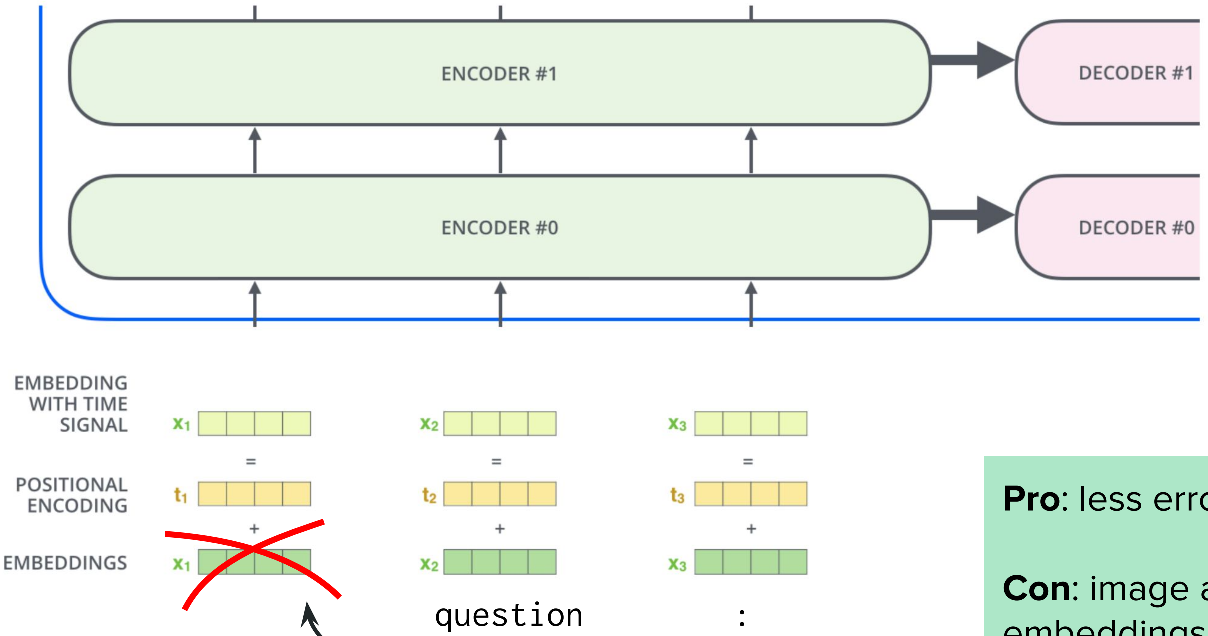
Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

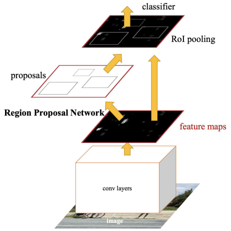
Hybrid fusion: Prepend **object vectors** to text embeddings



Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Hybrid fusion: Prepend **object vectors** to text embeddings





box feature vector ➞ project
box's coordinates ➞ project
sum

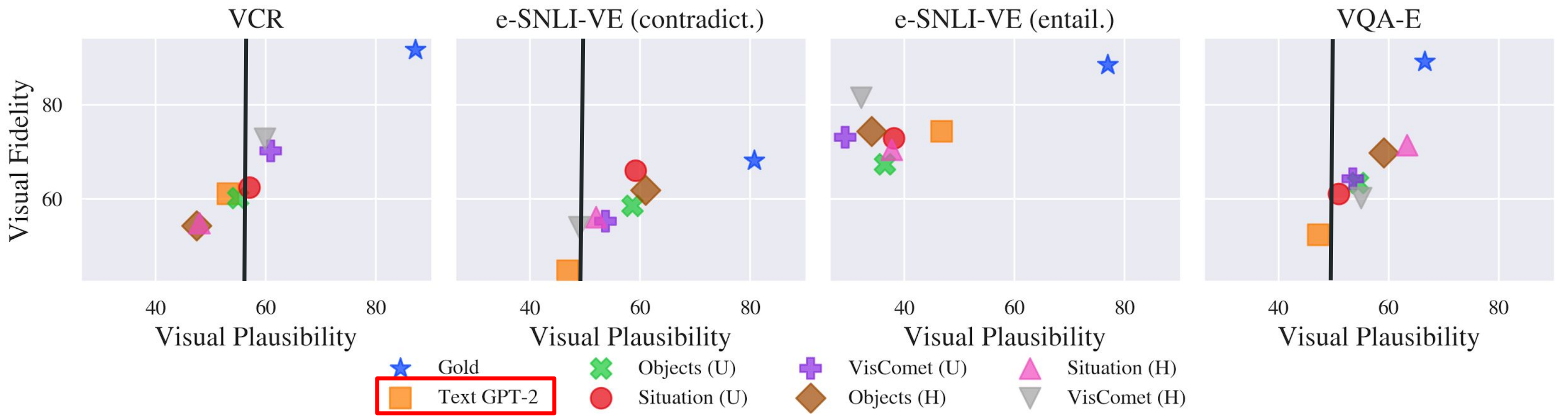
text-related features

Pro: less error-prone

Con: image and text embeddings come from different vector spaces

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

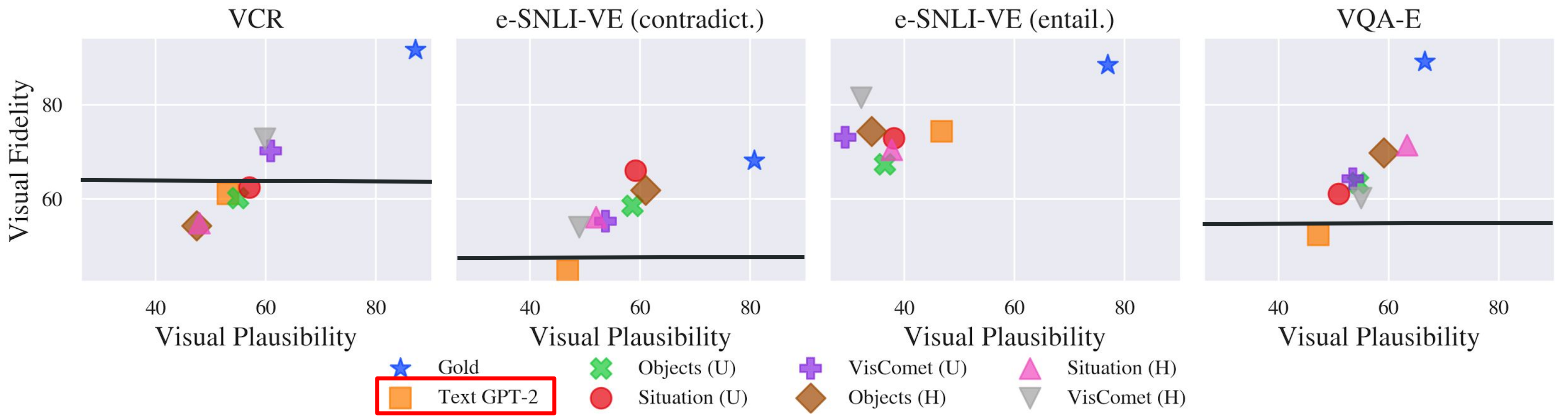
Summary of Results



➔ GPT-2 benefits from some form of visual adaptation for visual commonsense reasoning, visual-textual entailment, and visual question answering

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Summary of Results

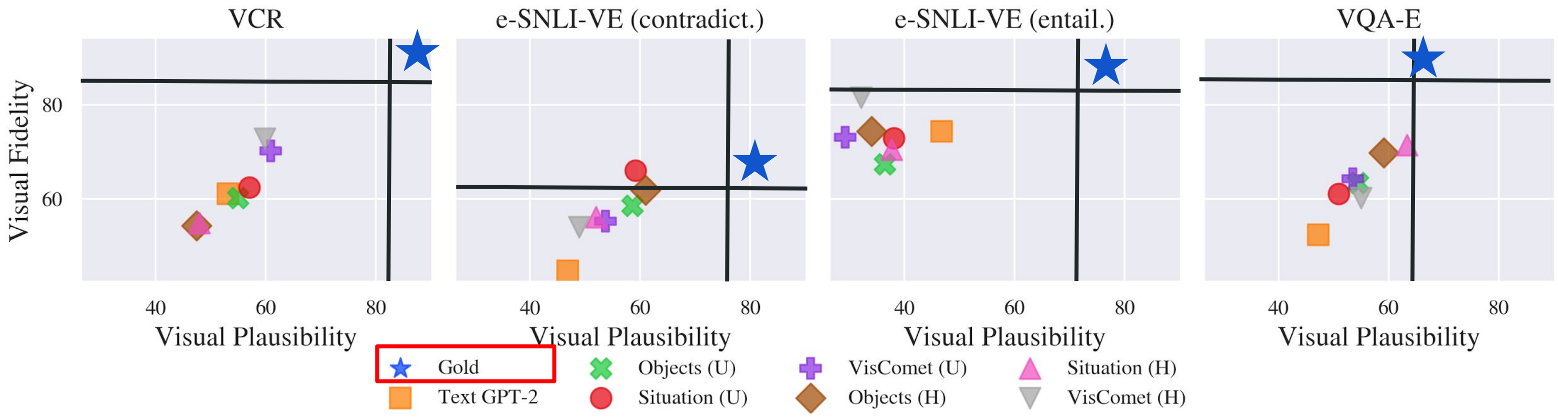


➔ GPT-2 benefits from some form of visual adaptation for visual commonsense reasoning, visual-textual entailment, and visual question answering

➔ Adapted models are less likely to mention content irrelevant to an image

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Summary of Results



➔ GPT-2 benefits from some form of visual adaptation for visual commonsense reasoning, visual-textual entailment, and visual question answering

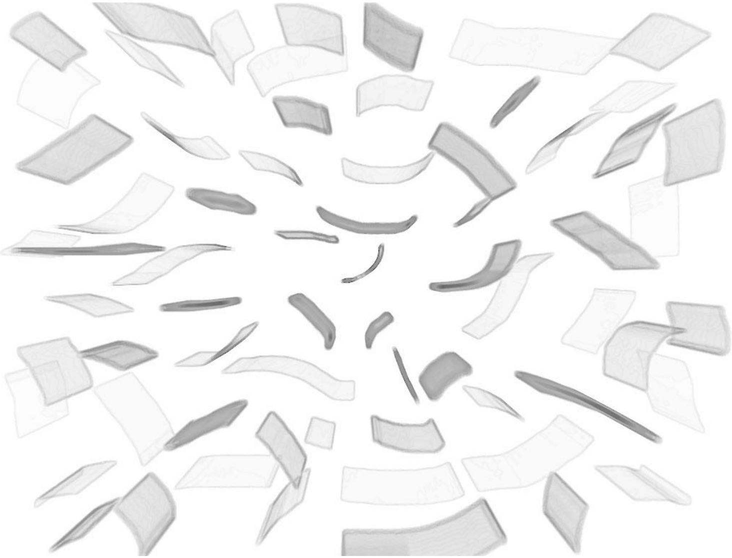
➔ Adapted models are less likely to mention content irrelevant to an image

➔ Best performing models are still behind human-written rationales

Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.

Future Direction:

Which developments of (multimodal) transformers are beneficial for the complex task of generating free-text explanations?



Future Direction:

Which developments of (multimodal) transformers are beneficial for the complex task of generating free-text explanations?

generation-suitable multimodal transformers[◊○]

tight vision and language

visually adapting language transformers (our work)

trained to generate complex text (implicitly capture some commonsense & world “knowledge”)

[◊] Zhou et al. Unified Vision-Language Pre-Training for Image Captioning and VQA. AAAI 2020.
[○] Gupta et al. Towards General Purpose Vision Systems.

Future Direction:

Which developments of (multimodal) transformers are beneficial for the complex task of generating free-text explanations?

generation-suitable multimodal transformers[◊O]

tight vision and language

visually adapting language transformers (our work)

trained to generate complex text (implicitly capture some commonsense & world "knowledge")

What is more important?

[◊] Zhou et al. Unified Vision-Language Pre-Training for Image Captioning and VQA. AAAI 2020.
[O] Gupta et al. Towards General Purpose Vision Systems.

Future Direction:

Which developments of (multimodal) transformers are beneficial for the complex task of generating free-text explanations?

generation-suitable multimodal transformers[◊O]

tight vision and language

visually adapting language transformers (our work)

trained to generate complex text (implicitly capture some commonsense & world "knowledge")

What is more important?

Does this depend on the finetuning data size?

[◊] Zhou et al. Unified Vision-Language Pre-Training for Image Captioning and VQA. AAAI 2020.
[O] Gupta et al. Towards General Purpose Vision Systems.

Future Direction:

Which developments of (multimodal) transformers are beneficial for the complex task of generating free-text explanations?

generation-suitable multimodal transformers[◊O]

tight vision and language

visually adapting language transformers (our work)

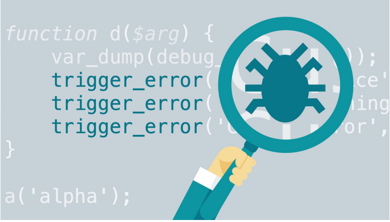
trained to generate complex text (implicitly capture some commonsense & world "knowledge")

What is more important?

Does this depend on the finetuning data size?

How about model size?

[◊] Zhou et al. Unified Vision-Language Pre-Training for Image Captioning and VQA. AAAI 2020.
[O] Gupta et al. Towards General Purpose Vision Systems.

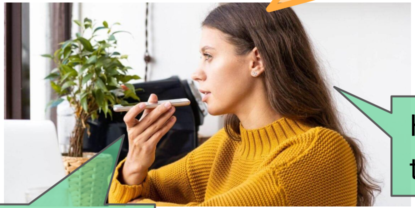


Technically robust and safe

Why this answer?[◇]

Allows acknowledging and evaluating trade-offs

Respects quality and integrity of data



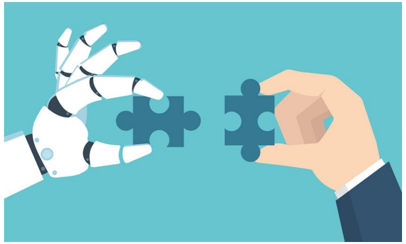
How to change the answer?[O]

Encourages green AI

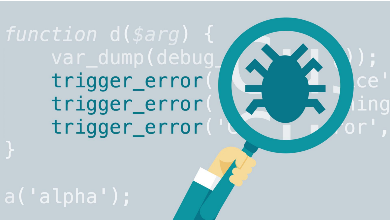
What if I change the input in this way?

Supports users' agency and oversight

Allows assessing the impact on individuals, society, democracy



European ethics guidelines for trustworthy AI
Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.
[◇] **Marasović** et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.
[O] Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

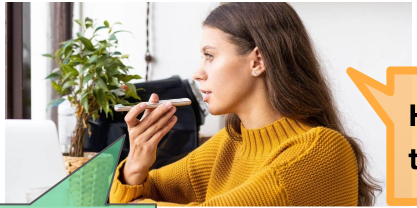


Technically robust and safe

Respects quality and integrity of data

Allows acknowledging and evaluating trade-offs

Why this answer?◇

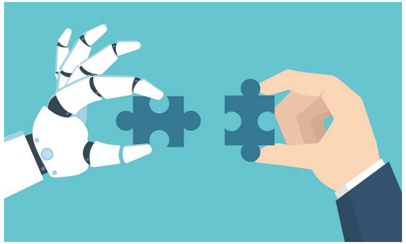


How to change the answer?○

What if I change the input in this way?

Encourages green AI

Supports users' agency and oversight



Allows assessing the impact on individuals, society, democracy

European ethics guidelines for trustworthy AI
Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.
◇**Marasović** et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. Findings of EMNLP 2020.
○ Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Miller's 2nd Insight from Social Science

Explanations are **contrastive** = responses to:

"Why P rather than Q?"

"How to change the answer from P to Q?"

where **P** is an observed event **(fact)**, and **Q** an imagined, counterfactual event that did not occur **(foil)**

Why did she get 20x less limit?

1. Make joint tax returns
2. Live in a community-property state
3. Be married for a long time
4. ....

What are the factors in the application that would need to change to get the same limit?

woman → ?

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

NLP is starting to pay attention!

COLING 2020 → Yang et al. Generating Plausible **Counterfactual Explanations** for Deep Transformers in Financial Text Classification.

TACL 2021 → Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution.

(Findings of) ACL 2021

→ Chen et al. KACE: Generating Knowledge-Aware **Contrastive Explanations** for NLI.

→ Ross et al. **Explaining** NLP Models via Minimal **Contrastive** Editing (MiCE).

→ Paranjape et al. Prompting **Contrastive Explanations** for Commonsense Reasoning Tasks.

→ Wu et al. Polyjuice: Generating **Counterfactuals** for **Explaining**, Evaluating, and Improving Models

EMNLP 2021 → Jacovi et al. **Contrastive Explanations** for Model Interpretability.

✅ Almost all of these papers begin by citing Miller’s overview of frameworks of explanations from social science

Are technical proposals the same?

Contrastive Explanations of NLP Models

Contrastive input editing:
Automatic edits to the input that change model output to the contrast case

Yang et al. COLING 2020.

Jacovi and Goldberg. TACL 2021.

Ross et al. Findings of ACL 2021.

Wu et al. ACL 2021.

Collect **free-text** human **contrastive explanations**, …

…and **generate them** left-to-right Chen et al. ACL 2021.

…abstract them into templates, automatically fill in the templates **(template-based infilling)**

Paranjape et al. Findings of ACL 2021.

Contrastive vector representation:
A dense representation of the input that captures latent features that differentiate two classes

Jacovi et al. EMNLP 2021.

Deeper Into Contrastive Editing

Contrastive Explanations via **Contrastive Editing**

The key idea:

“Why P not Q?” ⇒ “How to change the answer from P to Q?”

⇒ By making a **contrastive minimal edit**

A minimal edit to the input that causes the model output to change to the contrast case **has hallmark characteristics of a human contrastive explanation**:

→ cites contrastive features

→ selects a few relevant causes

Contrastive Explanations via **Contrastive Editing**

Question:
Ann and her children are going to Linda's home _____.

(a) by bus (b) by car (c) on foot (d) by train

Why **"by train"** (d) and not "**on foot"** (c)?
How to change the answer from **"by train"** (d) to "**on foot"** (c)?

Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at the train station. Our town is small...

MiCE-Edited Context:
...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at ~~the train station~~ **your home on foot**. Our ~~town~~ **house** is small...

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Alexis Ross, Ana Marasović, Matt Peters (2021)

Explaining NLP Models via Minimal Contrastive Editing (MiCE)

Goal:

Automatically find a **minimal edit** to the input that **causes the model output to change to the contrast case**

A very high-level idea of 🐭:

Keep masking and filling masked positions until you find an edit that flips the label, while simultaneously minimizing the masking percentage (i.e., the edit size)

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

mask *n*% of input tokens

```
label: positive input: Sylvester Stallone has made some <mask> films in his lifetime,
but this has got to be one of the <mask>. A totally <mask> story...
```

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

mask *n*% of input tokens

```
label: positive input: Sylvester Stallone has made some <mask> films in his lifetime,
but this has got to be one of the <mask>. A totally <mask> story...
```

sample *m* spans at each masked position

```
1.   label: positive input: Sylvester Stallone has made some good films in his lifetime,
     but this has got to be one of the worst. A totally novel story...

2.   label: positive input: Sylvester Stallone has made some great films in his lifetime,
     but this has got to be one of the greatest of all time. A totally boring story...

...

m.  label: positive input: Sylvester Stallone has made some wonderful films in
    his lifetime, but this has got to be one of the greatest. A totally tedious
    story...
```

the contrast label (foil)

```
label: positive input: Sylvester Stallone has made some crap films in his lifetime,
but this has got to be one of the worst. A totally dull story...
```

mask *n*% of input tokens

```
label: positive input: Sylvester Stallone has made some <mask> films in his lifetime,
but this has got to be one of the <mask>. A totally <mask> story...
```

sample *m* spans at each masked position

get the probability of the contrast label

```
1.   label: positive input: Sylvester Stallone has made some good films in his lifetime,
     but this has got to be one of the worst. A totally novel story...
```

$\mathbb{P}(pos) = 0.2$

```
2.   label: positive input: Sylvester Stallone has made some great films in his lifetime,
     but this has got to be one of the greatest of all time. A totally boring story...
```

$\mathbb{P}(pos) = 0.6$

...

```
m.  label: positive input: Sylvester Stallone has made some wonderful films in
    his lifetime, but this has got to be one of the greatest. A totally tedious
    story...
```

$\mathbb{P}(pos) = 0.65$

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

×

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

➔ If a contrastive edit found: $n^{(2)}$=13.75%

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

→ If a contrastive edit found: $n^{(2)}$=13.75%

→ If a contrastive edit **not** found: $n^{(2)}$=41.25%

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}$=27.5%

- ➔ If a contrastive edit found: $n^{(2)}$=13.75%
  - ◆ If a contrastive edit found: $n^{(3)}$=6.875%
- ➔ If a contrastive edit **not** found: $n^{(2)}$=41.25%
  - ◆ If a contrastive edit found: $n^{(3)}$=20.625%

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

$\times$

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick which values for *n*?

Binary search on [0,55]

Start: $n^{(1)}=27.5\%$

- ➔ If a contrastive edit found: $n^{(2)}=13.75\%$
    - If a contrastive edit found: $n^{(3)}=6.875\%$
    - If a contrastive edit **not** found: $n^{(3)}=20.625\%$
- ➔ If a contrastive edit **not** found: $n^{(2)}=41.25\%$
    - If a contrastive edit found: $n^{(3)}=20.625\%$
    - If a contrastive edit **not** found: $n^{(3)}=48.125\%$

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

×

s* different values of *n* to minimize the edit

* s=4 in the paper

How to pick masking positions?

Based on token importance for the original prediction

Rank input tokens based on the magnitude of the gradients of the model we're explaining

Mask top-*n*% of **ranked** tokens

We find that this works better than randomly masking tokens

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

×

s* different values of *n* to minimize the edit

* s=4 in the paper

*s***m* samples

* m=15 in the paper

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample m spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

1. Prepend the contrast label to the input
2. Mask n% of the input tokens
3. Sample m spans at masked positions

× **s different values of n to minimize the edit***

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

beam keep top-b samples

* b=3 in the paper

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

1. Prepend the contrast label to the input
2. Mask *n%* of the input tokens
3. Sample *m* spans at masked positions

× ***s* different values of *n* to minimize the edit***

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

beam keep top-*b* samples

* b=3 in the paper

if the contrastive edit is found

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

repeat these steps for every instance in the beam for 2 more rounds

1. Prepend the contrast label to the input
2. Mask *n*% of the input tokens
3. Sample *m* spans at masked positions

× **_s_ different values of _n_ to minimize the edit***

* s=4 in the paper

$s*m$ samples

* m=15 in the paper

rank $s*m$ samples w.r.t. the probability of the contrast label

beam keep top-*b* samples

* b=3 in the paper

The maximum number of iterations for a single instance:

$$\overbrace{\#\text{ binary search levels } \mathbf{s} \times \#\text{ samples at each maskin position } \mathtt{m}}^{\text{first round}} +$$

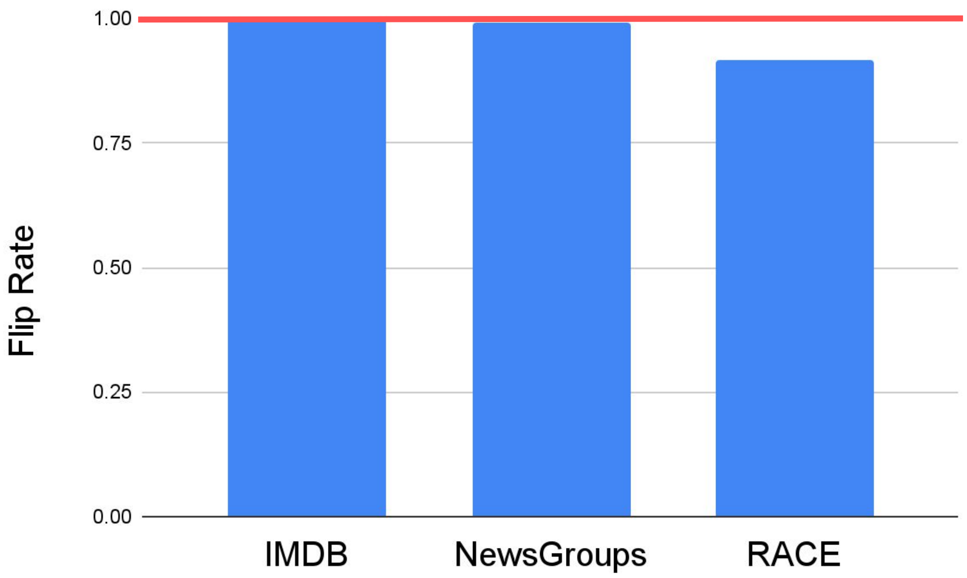
$$\underbrace{\text{beam size } \mathtt{b} \times \#\text{ binary search levels } \mathbf{s} \times \#\text{ samples at each masking position } \mathtt{m} \times \#\text{ of rounds}}_{\text{other rounds}} =$$

$4 \times 15 + 3 \times 4 \times 15 \times 2 = 420$

(that’s a lot)

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Results – Flip Rate

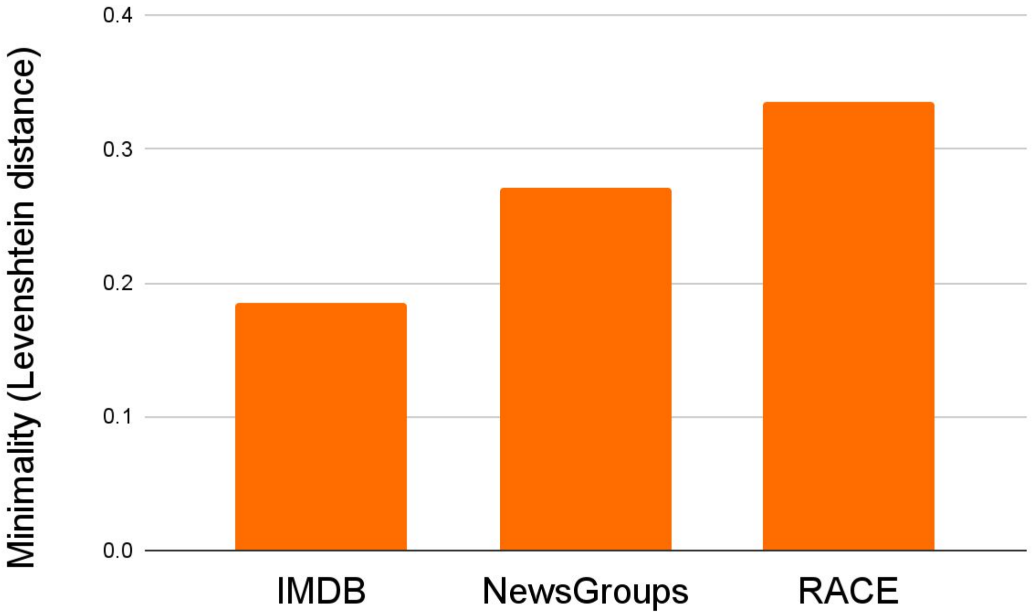


1.0 when we find a contrastive edit for all instances

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Results – Edit Minimality

The minimum number of deletions, insertions, or substitutions required to transform the original to the
edited instance

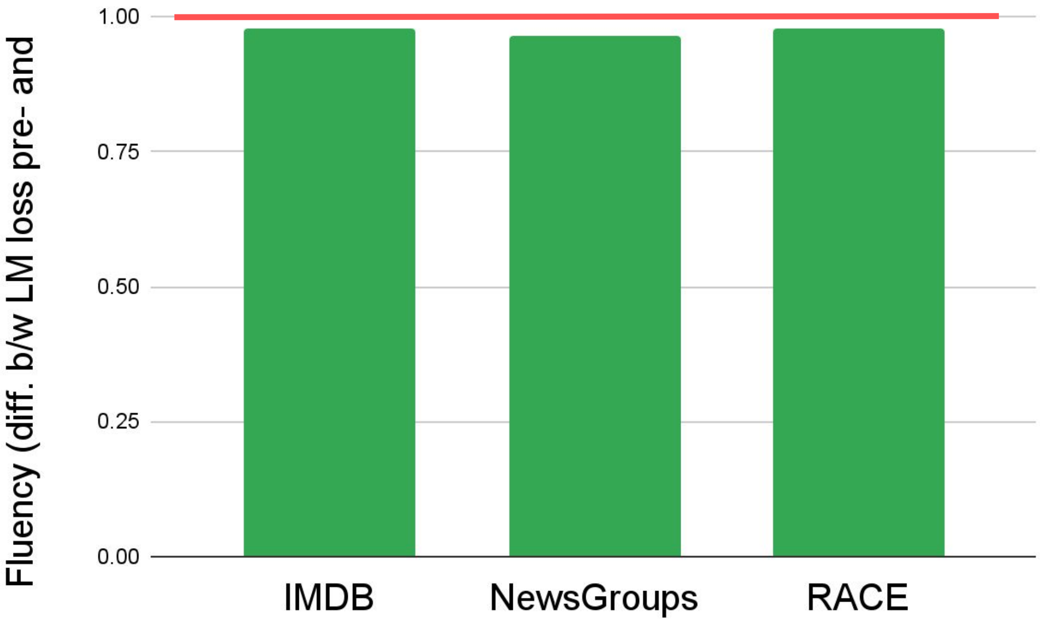


lower is better; we change on average 18.5-33.5% of the input tokens

The size of the IMDB edits is similar to human edits*

* Compared to IMDB edits in Gardner et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. Findings of EMNLP 2020.

Results – Edit Fluency



1.0 when a LM loss pre- and post-editing doesn’t change

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

How Can MiCE Edits Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Original pred $y_p =$ positive **Contrast pred** $y_c =$ negative

An interesting pairing of stories, this little flick manages to bring together seemingly different characters and story lines all in the backdrop of WWII and succeeds in tying them together without losing the audience. I was impressed by the depth portrayed by the different characters and also by how much I really felt I understood them and their motivations, even though the time spent on the development of each character was very limited. The outstanding acting abilities of the individuals involved with this picture are easily noted. A fun, stylized movie with a slew of comic moments and a bunch more head shaking events. ~~7/10~~ **4/10**

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

How Can MiCE Edits Be Used?

MiCE's edits can offer hypotheses about model "bugs"

Hypothesis:
Model learned to rely heavily on numerical ratings ⭐

Test the hypothesis using MiCE's edits:

1. Filter instances for which the MiCE edit has a minimality value of ≤ 0.05

2. Select tokens that are removed/inserted at a higher rate than expected given the frequency with which they appear in the original IMDB inputs

$y_c =$ ***positive***		$y_c =$ ***negative***	
Removed	**Inserted**	**Removed**	**Inserted**
4/10	excellent	10/10	awful
ridiculous	enjoy	8/10	disappointed
horrible	amazing	7/10	1
4	entertaining	9	4
predictable	10	enjoyable	annoying

✅ NLP is starting to acknowledge the perspective of the social sciences on explainability

✅ Contrastive editing is already achieving decent performance

❗ Obviously needed improvements:

- ➔ less iterations
- ➔ more precise minimality

(Contrastive) Local Explanations: What is Next?

Miller's 1st Insight from Social Science

Explanation are **selected (in a biased manner)** because:

1. **Cognitive load**: causal chains are often too large to comprehend

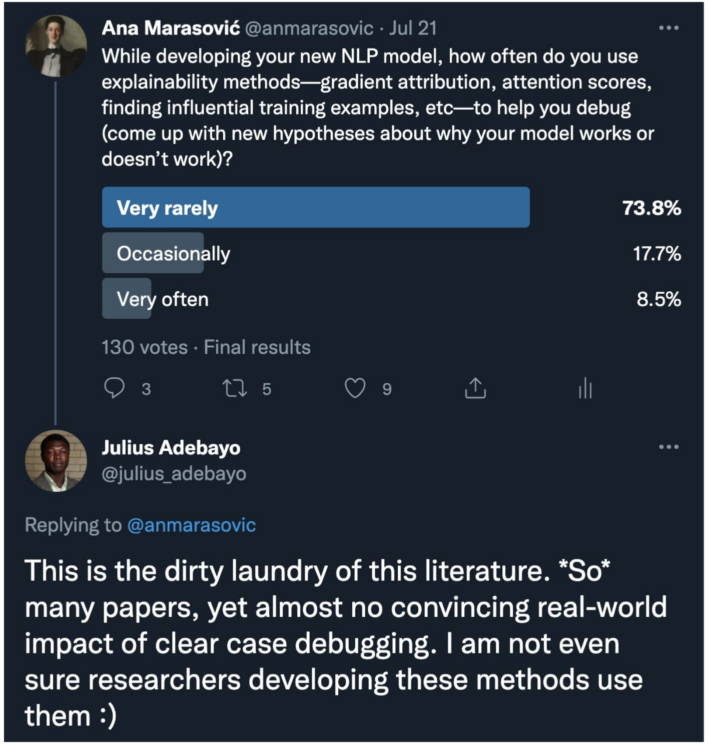
2. Explainee cares only about a small number of causes (relevant to the context)

We don't test whether generated contrastive explanations are more easily understood or whether they match people's expectations

Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019.

This is not specific to contrastive explanations...

Although local explanations are specifically motivated for people to use, there is no convincing evidence that local explanations help people who are using language technology



We Lack Evidence That Local Explanations Are Helpful

This is in part due to:

- **Focus on grand AI challenges**, but not useful applications
- **Simple tasks** that people don't need help with (e.g., commonsense QA)
- The use of automatic measures of explanation plausibility **without specifying what real-world situations highly plausible explanations will help with**

Future Direction:

How and when are local explanations useful?

This is in part due to:

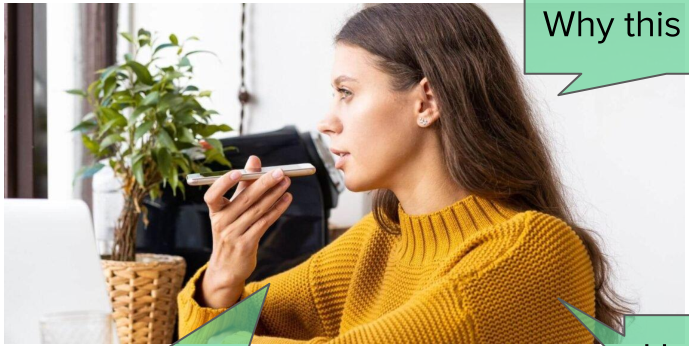
- **Focus on grand AI challenges**, but not useful applications
- **Simple tasks** that people don't need help with (e.g., commonsense QA)
- The use of automatic measures of explanation plausibility **without specifying what real-world situations highly plausible explanations will help with**

To meaningfully move forward we need to answer:

➔ **What are potentially useful language applications and who is targeted audience?**
(e.g., journalist and fact checking)

➔ **How explanations might help people using these applications?**
(e.g., by helping them verify information *faster* without the loss of *accuracy*)

➔ **Test them exactly for those purposes**



Why this answer?

What if I change the input in this way?

How to change the answer?

Data NeurIPS 2021

Modeling EMNLP 2020

Theoretical and Empirical Evaluation
EMNLP 2021
Findings of ACL 2021
FAccT 2021

Findings of ACL 2021 🐭

What if I encode the linguistic structure differently?

Hoyle, **Marasović**, Smith. Promoting Graph Awareness in Linearized Graph-to-Text Generation. Findings of ACL 2021.

What if I change the data domain?

Gururangan, **Marasović**, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

Marasović and Frank. Multilingual Modal Sense Classification using a Convolutional Neural Network. Repl4NLP 2016.

Marasović et al. Modal Sense Classification At Large. LiLT 2016.

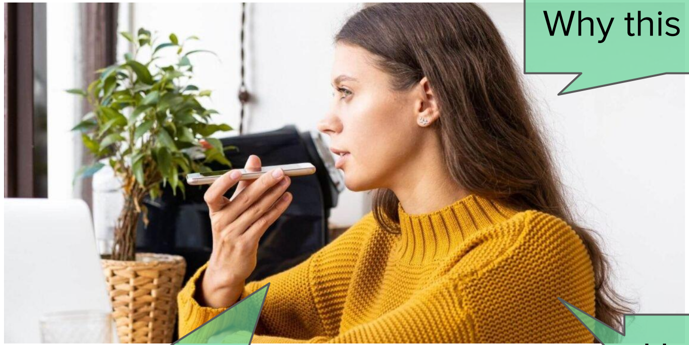
What if a certain language phenomenon is present?

Dasigi, Liu, **Marasović**, Smith, Gardner. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. EMNLP 2019.

What if the training data is limited?

Marasović and Frank. Improving Opinion Role Labeling Using Multi-Task Learning with Semantic Role Labeling. NAACL 2018.

Marasović et al. A Mention-Ranking Model for Abstract Anaphora Resolution. EMNLP 2017.



Why this answer?

What if I change the input in this way?

How to change the answer?

Data NeurIPS 2021

Modeling Findings of EMNLP 2020

Theoretical and Empirical Evaluation
EMNLP 2021
Findings of ACL 2021
FAccT 2021

Findings of ACL 2021 🐭

What if I encode the linguistic structure differently?
Findings of ACL 2021

What if I change the data domain?
ACL 2020
Repl4NLP 2016
LiLT 2016

What if a certain language phenomenon is present?
EMNLP 2019

What if the training data is limited?
NAACL 2018
EMNLP 2017

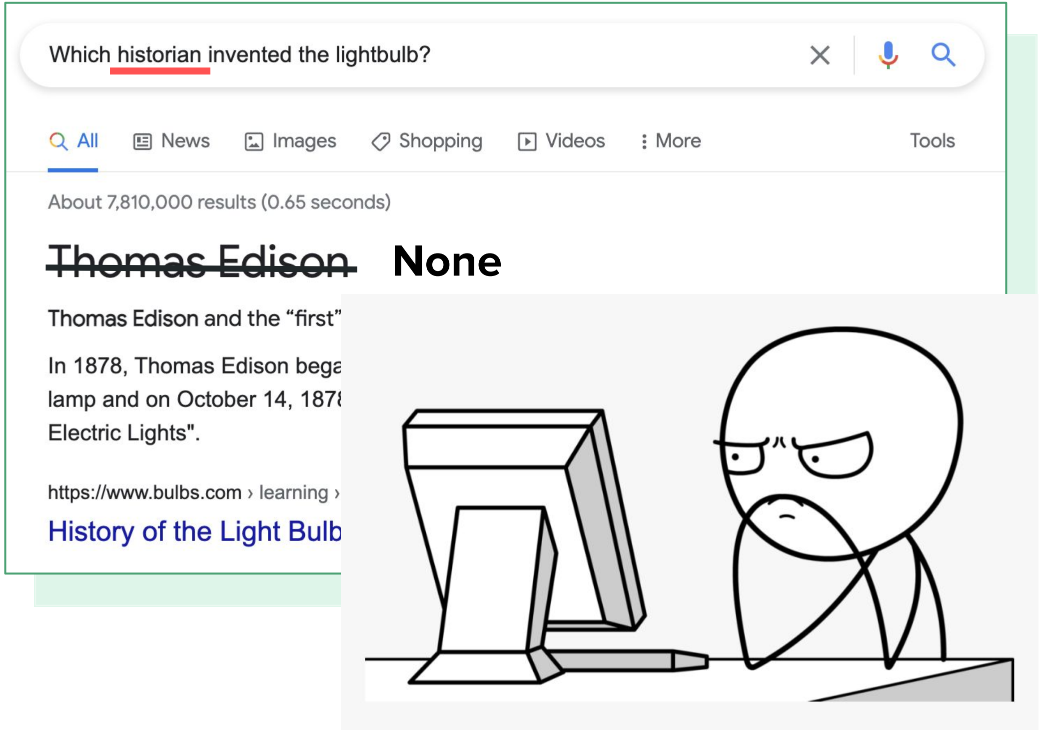
Why this answer?

What if I change the input in this way?

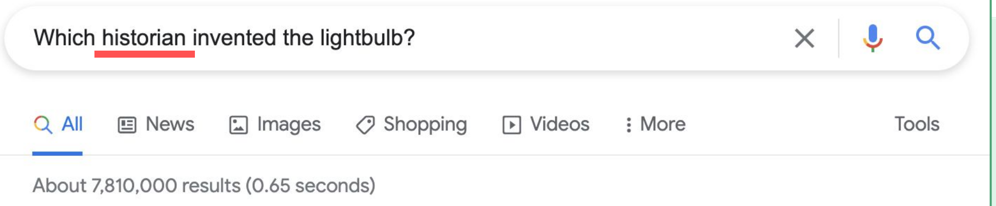
How to change the answer?

THANK YOU!
YOUR QUESTION ⤵

• • •



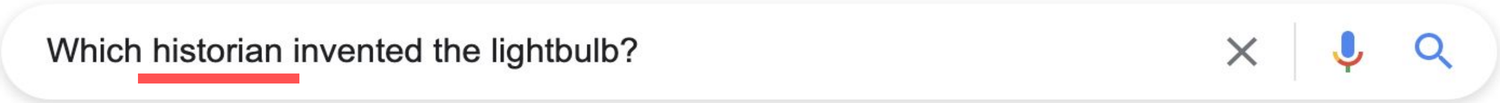
Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.



~~Thomas Edison~~

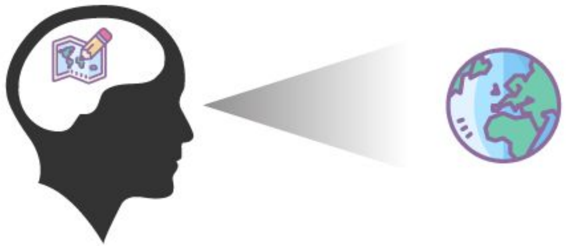
None *because* Thomas Edison is credited as the primary inventor of the lightbulb and Edison was not a historian

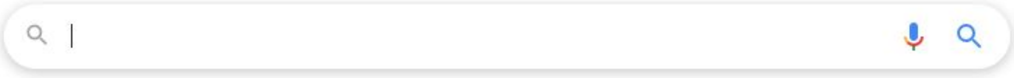
Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.



constrain the system to explain
"why is this input assigned this answer"
to be more intuitive to people

"None because Thomas Edison is credited as the primary inventor of the lightbulb and Edison was not a historian"

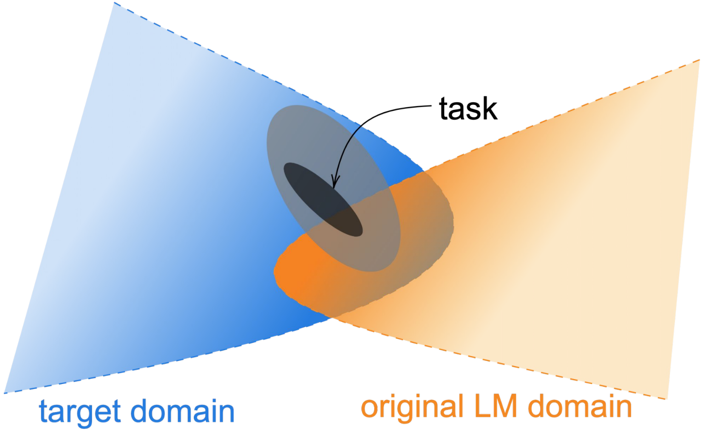


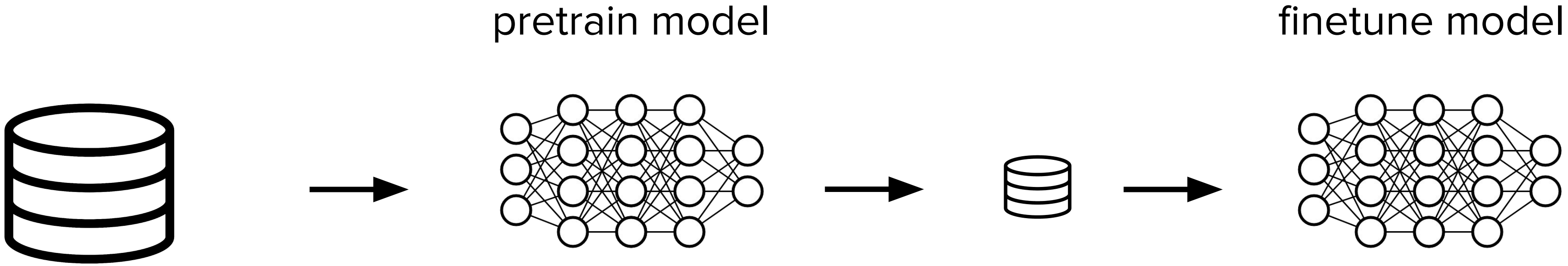


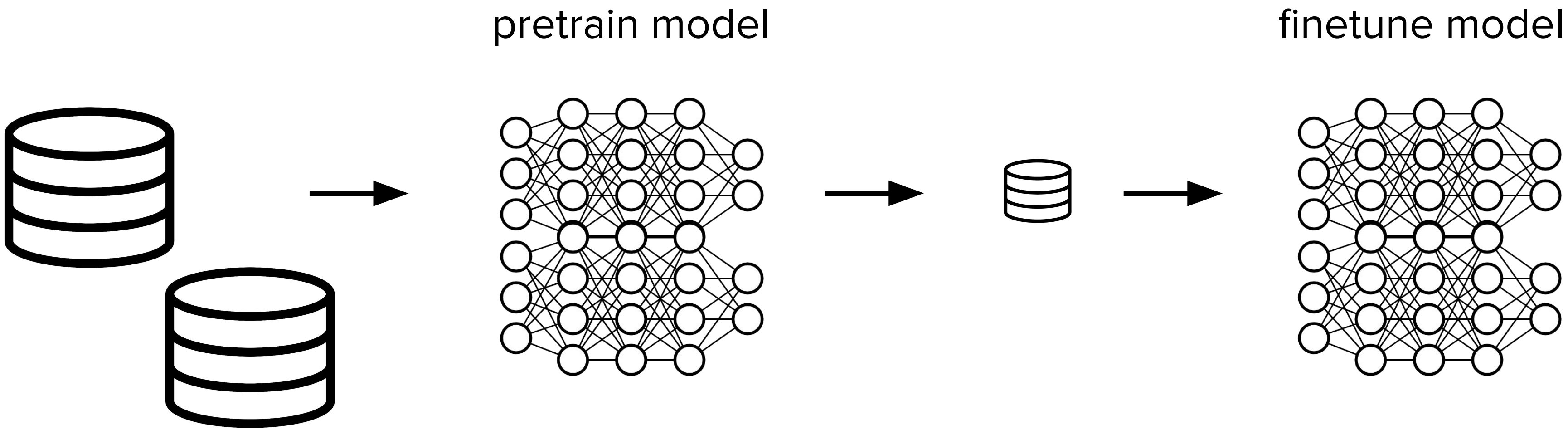
mental model about how to interact and control the system

Gururangan, Marasović, Swayamdipta, Lo, Beltagy, Downey, Smith (2020):

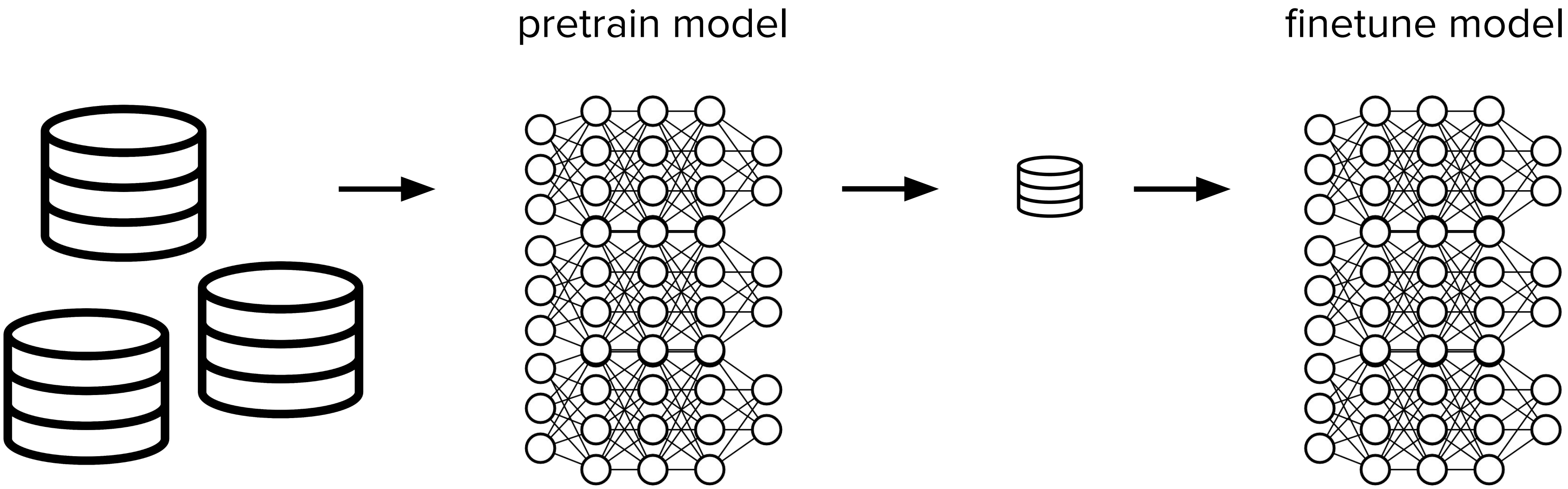
Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks

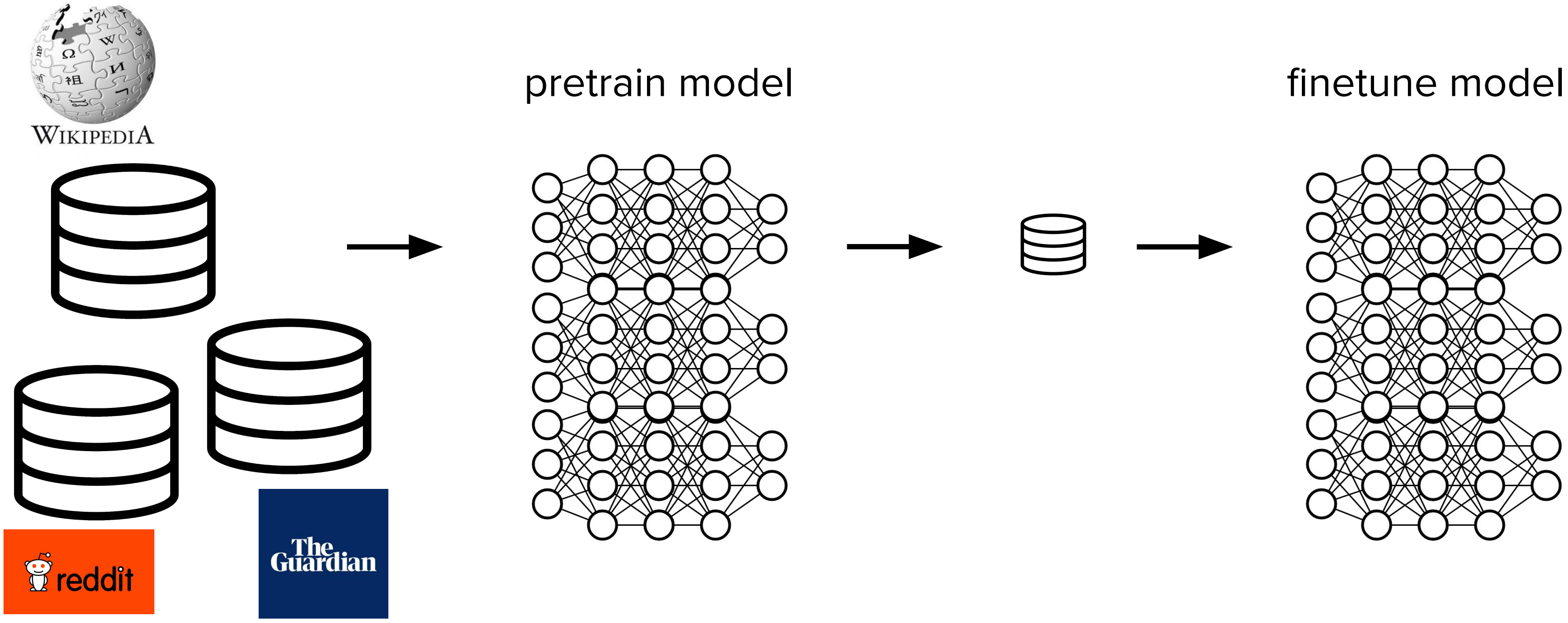



pretrain model
finetune model



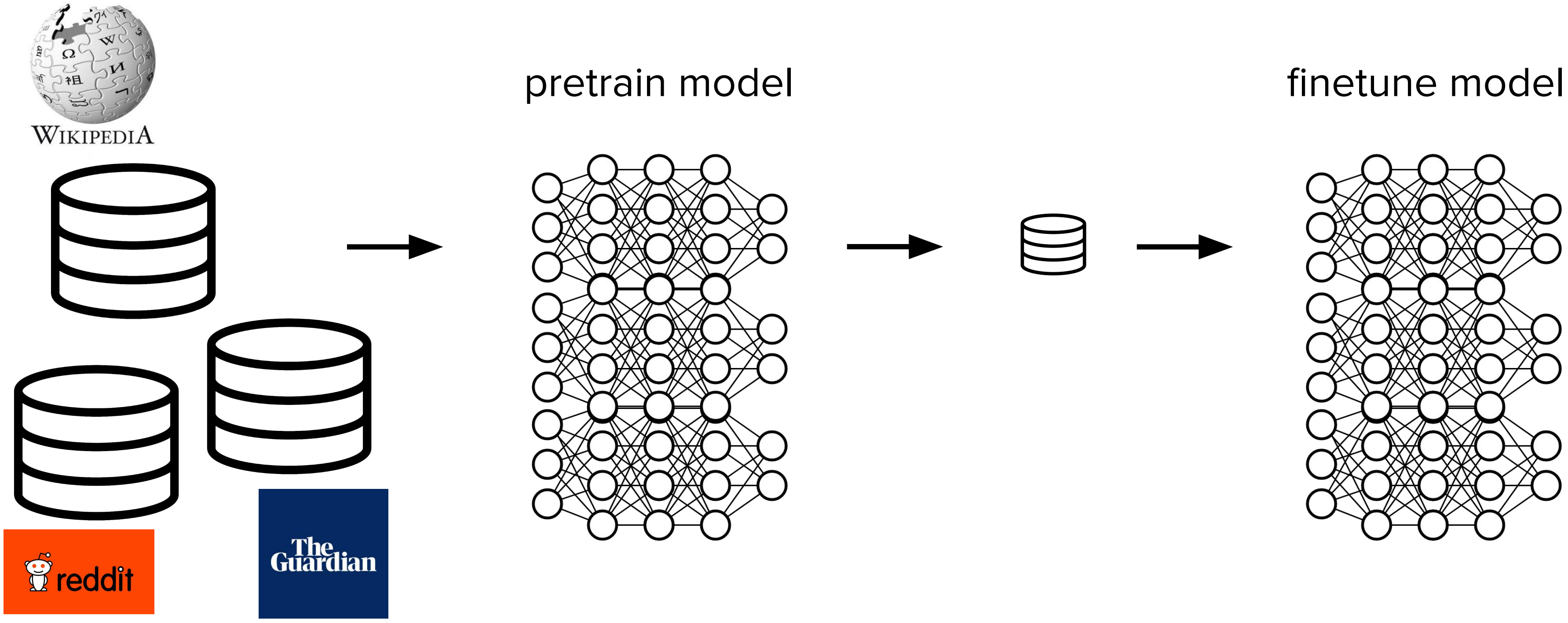
pretrain model
finetune model


pretrain model
finetune model



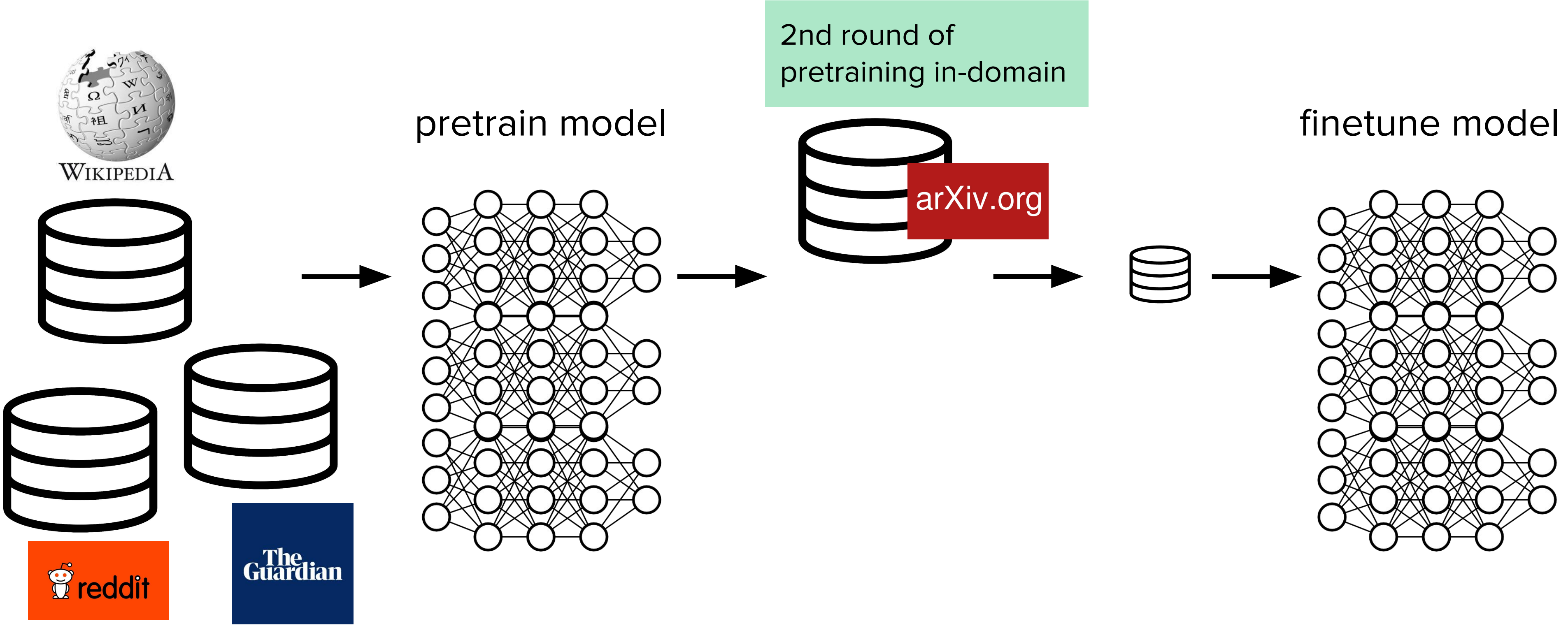
Dodge, Sap, **Marasović**, Agnew, Ilharco, Groeneveld, Mitchell, Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. EMNLP 2021.

What if I change the data domain? Do the latest large pretrained models work universally?



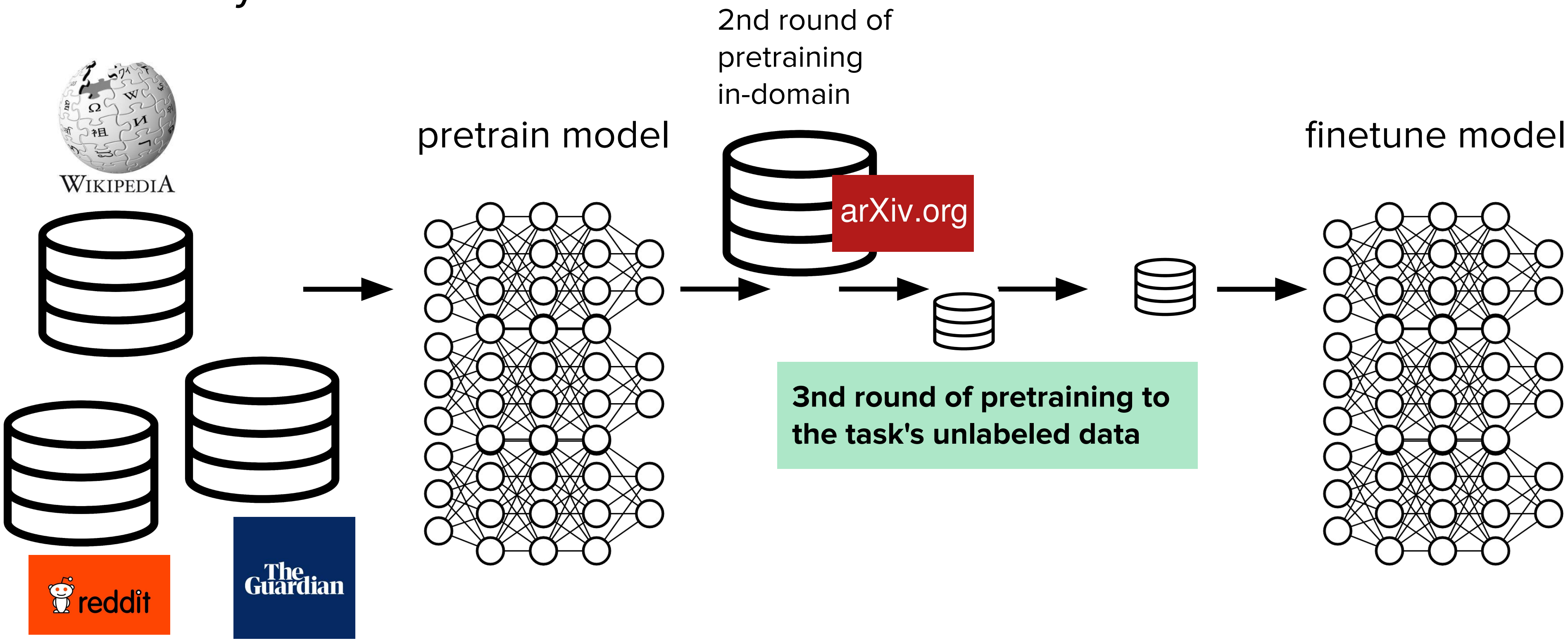
Gururangan, **Marasović**, Swayamdipta, Lo, Beltagy, Downey, Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

Summary of Results – Part II



Gururangan, **Marasović**, Swayamdipta, Lo, Beltagy, Downey, Smith. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

Summary of Results – Part II

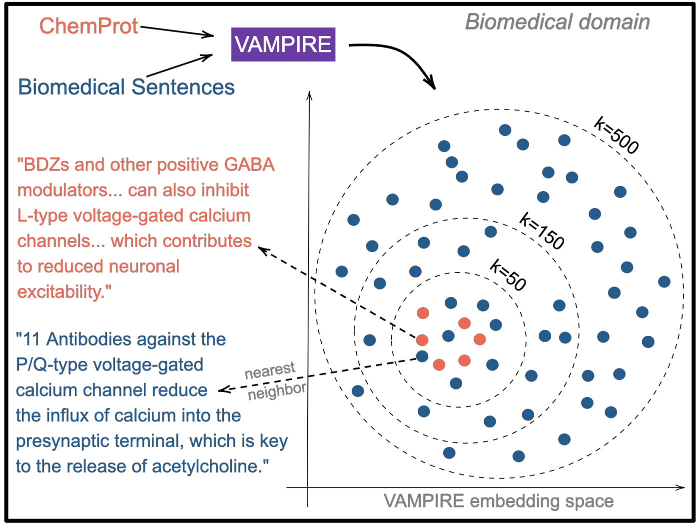


Gururangan, **Marasović**, Swayamdipta, Lo, Beltagy, Downey, Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ACL 2020.

Summary of Results – Part III

Adapting to a task corpus augmented using simple data selection strategies is an effective alternative

It may be valuable to complement work on ever-larger LMs with parallel efforts to identify and use domain- and task-relevant corpora to specialize models



VAMPIRE → Gururangan et al. Variational Pretraining for Semi-supervised Text Classification. ACL 2019.

Can a pretrained model without any additional tweaks fill in the spans?

So-so

We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it for editing

`(standard masking)` `Sylvester Stallone has made some` **`<mask>`** `films in his lifetime, but this has got to be one of the` **`<mask>`**`. A totally` **`<mask>`** `story...`

`(targeted masking)` `label: positive input: Sylvester Stallone has made some` **`<mask>`** `films in his lifetime, but this has got to be one of the` **`<mask>`**`. A totally` **`<mask>`** `story...`

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Can a pretrained model without any additional tweaks fill in the spans?

So-so

We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it for editing

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance (good option if you don't have the labeled data)

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.

Can a pretrained model without any additional tweaks fill in the spans?

So-so

We find that **preparing the editor** by finetuning it to infill masked spans given masked text and **a target end-task label** as input is an important step before using it for editing

We find that **labels predicted by the model** we're explaining **can be used** in this step without a big loss in performance (good option if you don't have the labeled data)

⇒ **MiCE is a two-stage approach** to generating contrastive edits

Stage 1: prepare an editor
Stage 2: makes edits guided with gradients & logits of the model we're explaining

Ross, **Marasović**, Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021.