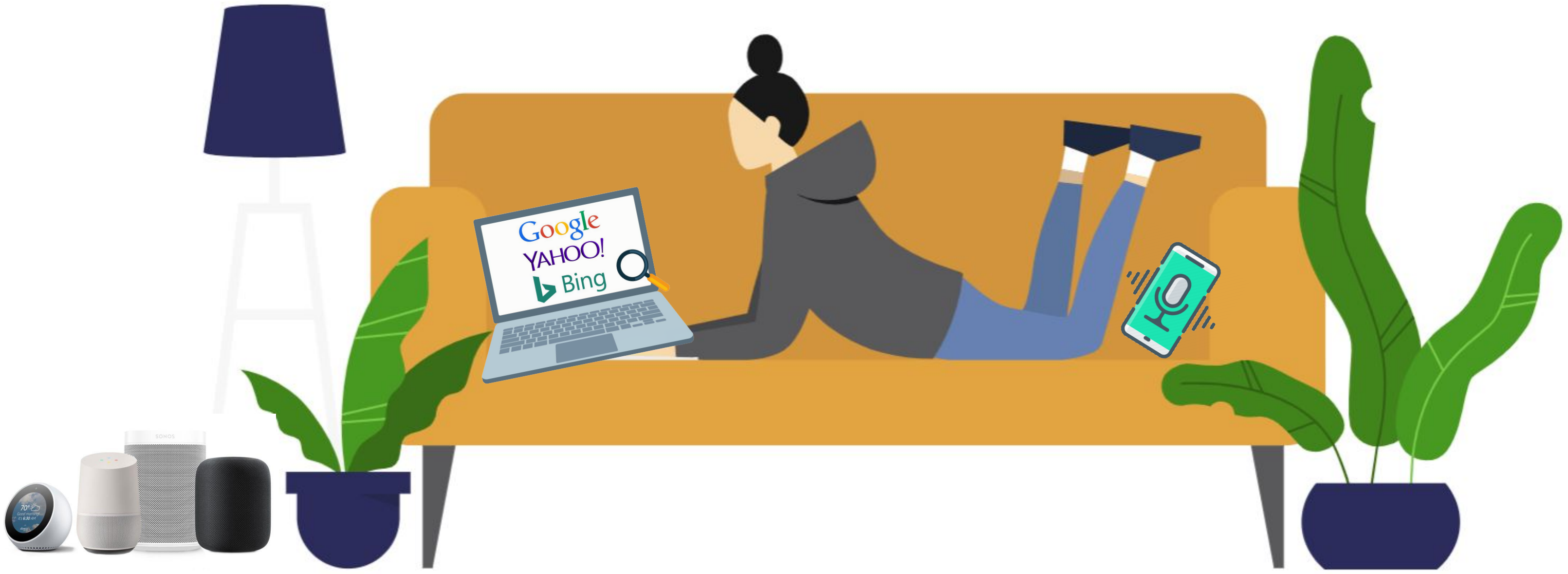
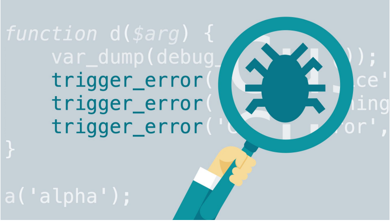
# Self-Explainability for Intuitive & Controllable Interaction: **On Reducing Human-Authored Free-Text Explanations for Training**

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

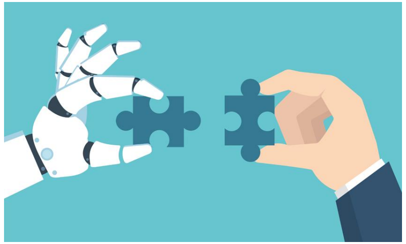
# Trustworthy AI Contracts

Respects quality and integrity of data

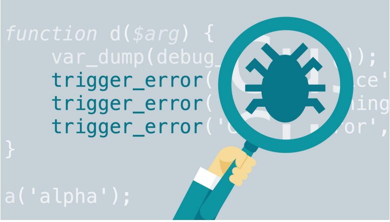
Encourages green AI

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.



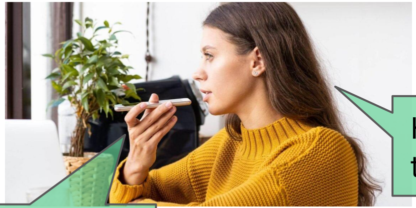

**Technically robust and safe**

**Respects quality and integrity of data**

**Allows acknowledging and evaluating trade-offs**

**Encourages green AI**

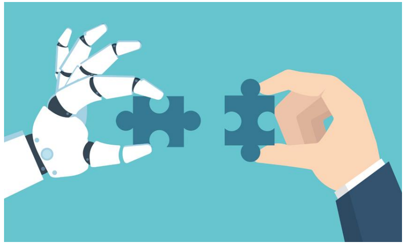
Why this answer?



How to change the answer?

What if I change the input in this way?

**Supports users' agency and oversight**

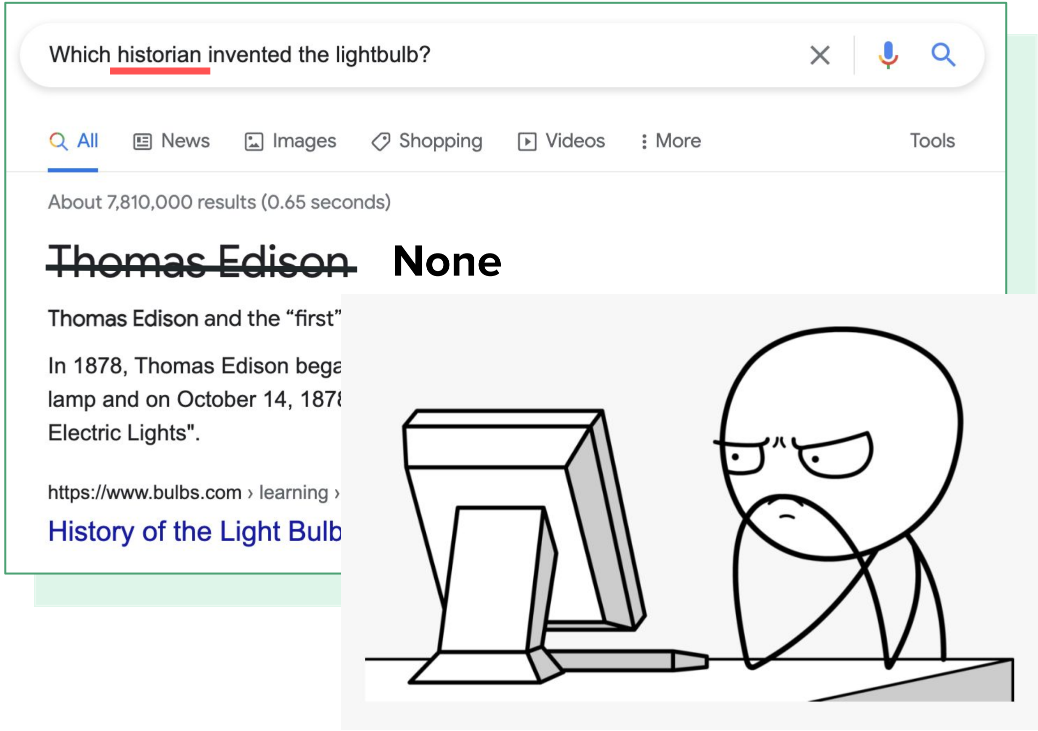


**Allows assessing the impact on individuals, society, democracy**

European ethics guidelines for trustworthy AI

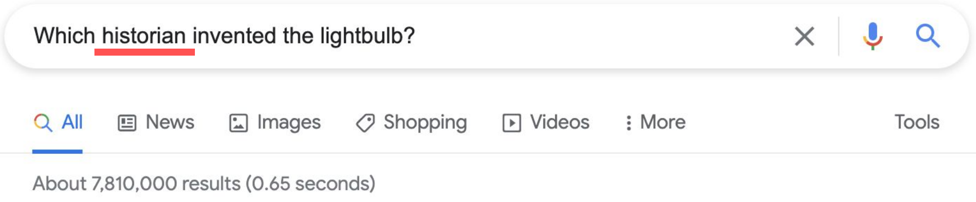
Jacovi, **Marasović**, Miller, Goldberg. Formalizing Trust in Artificial Intelligence. FAccT 2021.






**None**

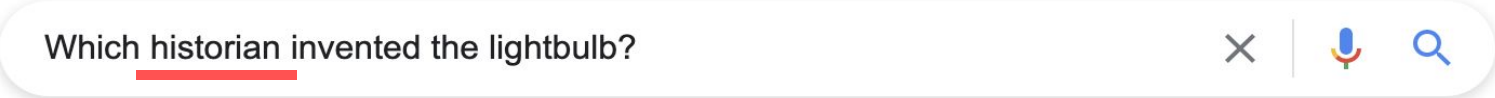
Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.


~~Thomas Edison~~

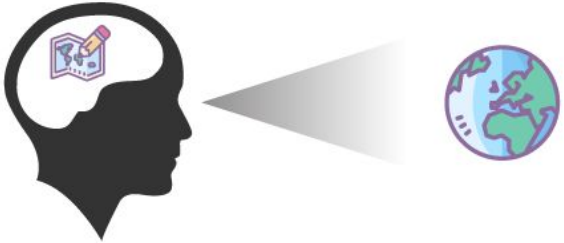
**None *because* Thomas Edison is credited as the primary inventor of the lightbulb and Edison was not a historian**

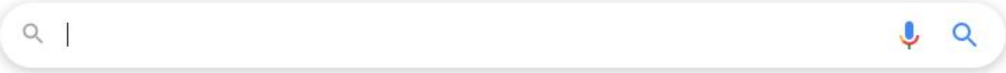
Example from: Kim et al. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2021.




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**

**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.***

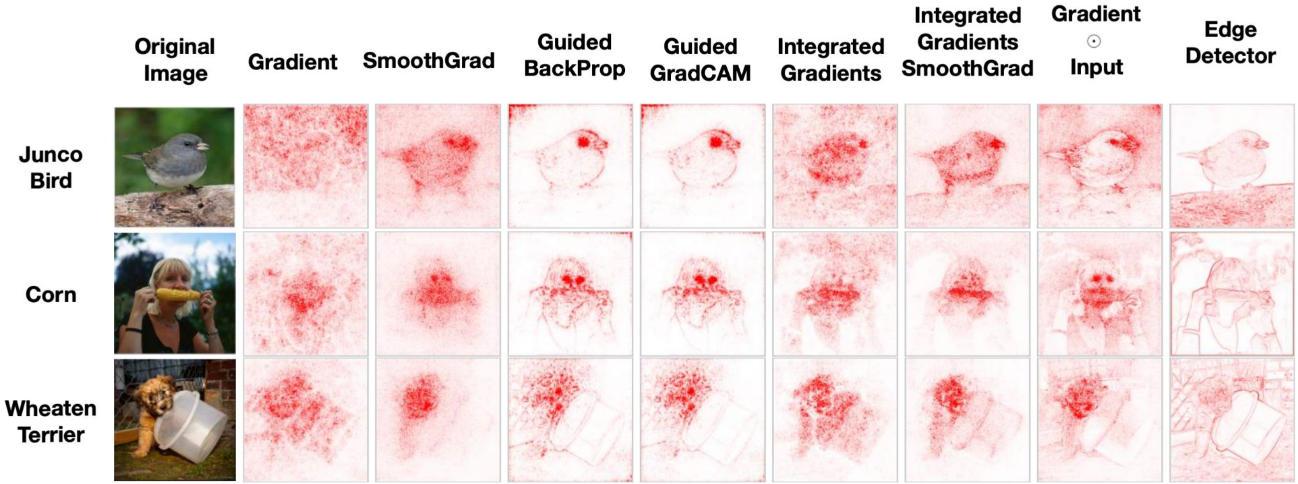
# Answering “why” by highlighting

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

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

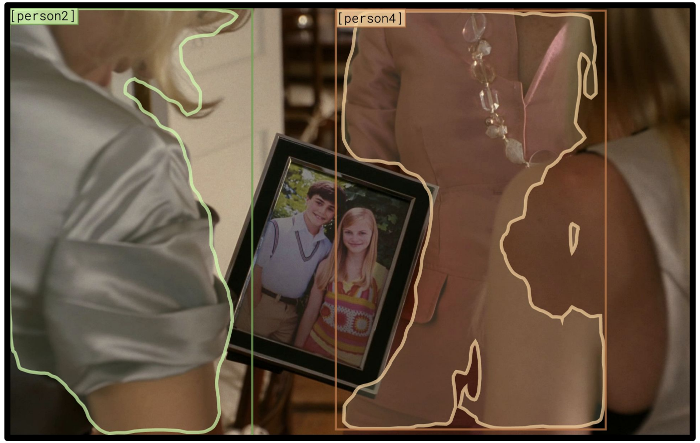
# Answering “why” by highlighting




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

Answering “why” by highlighting…

## …doesn’t work when the reason is not explicitly stated in the input




**Question**: What is going to happen next?

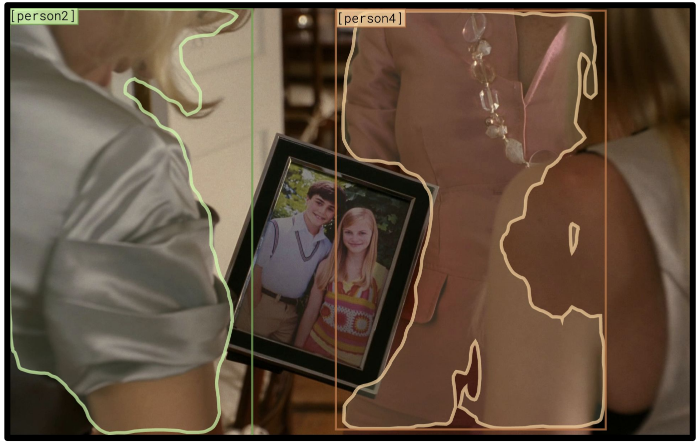
**Answer**: [person2] holding the photo will tell [person4] how cute their children are.

**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.

Answering “why” by highlighting...

## ...doesn’t work 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.

# That’s great, but...

Current self-rationalization models rely on an **abundance** of **human-written explanations** for **each task** (Narang et al., 2020)

# Everyone wants to minimize data annotation anyway

**Prompting**

- In-context learning (GPT-3 style)
- Prompt-based finetuning
- Automatic prompt search

**Supplementing LM pretraining**

- Domain- or task-specific unlabeled data (Gururangan et al., 2020)
- Automatically generating labeled data (Lewis et al., 2019)
- Human-annotated data of data-rich tasks (Phang et al, 2020)

***Beyond classification & "SQuAD" QA?***

# Today

**Long first part:**

**Prompt-based finetuning for self-rationalization**

**Brief second part:**

**Training with auto. extracted question-answer-explanation instances**

# Few-shot Self-rationalization

*Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

# Principles of prompt-based finetuning

A pretrained LM is well-positioned to solve the end-task if...

...we **format finetuning end-task examples as similar as possible to the format used in pretraining**

# Self-rationalization models…

…are currently T5-based* because:

- T5 has been pretrained on a mix of **span-filling** with various **supervised tasks** including classification, QA, and generation

- T5-variants are largest *available* pretrained LMs (11B)

* [Narang et al., 2020](); [Hase et al., 2020](); [Wiegreffe et al., 2021]()

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

## Natural Language Inference ([Camburu et al., 2018])

Premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement.

Hypothesis: It is empty outside.

Label: contradiction

Explanation: It can not be empty outside if people are standing outside in a crowd.

## Span infilling

`model_input:` explain nli hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement. <extra_id_0> because <extra_id_1>

`model_output:` <extra_id_0> contradiction <extra_id_1> it can not be empty outside if people are standing outside in a crowd.<extra_id_2>

T5 is pretrained with MNLI and infilling

## Span infilling

model_input: explain nli hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement. <extra_id_0> because <extra_id_1>

model_output: <extra_id_0> contradiction <extra_id_1> it can not be empty outside if people are standing outside in a crowd.<extra_id_2>

## T5’s NLI

model_input: explain nli hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement.

model_output: contradiction because it can not be empty outside if people are standing outside in a crowd.

# But we wish to rationalization any task

# But we wish to rationalization any task

**CommonsenseQA** (Aggarwal et al., 2021)

Question: Where is a frisbee in play likely to be?
Choices**:** outside, park, roof, tree, air
Answer: air 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.

# But we wish to rationalization any task

**CommonsenseQA** (Aggarwal et al., 2021)

Question: Where is a frisbee in play likely to be?
Choices**:** outside, park, roof, tree, air
Answer: air 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.

**SBIC** (Sap et al., 2020)

Post: Have you ever tasted Ethiopian food? You haven't? Don't worry they haven't either..

Label: offensive Explanation: This post implies that ethiopians are starving.

# But we wish to rationalization any task

**CommonsenseQA** (Aggarwal et al., 2021)

Question: Where is a frisbee in play likely to be?
Choices**:** outside, park, roof, tree, air
Answer: air 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.

**SBIC** (Sap et al., 2020)

Post: Have you ever tasted Ethiopian food? You haven't? Don't worry they haven't either..

Label: offensive Explanation: This post implies that ethiopians are starving.

**ComVE** (Wang et al., 2019)

Sentence1: The stove was cleaned with a cleaner.

Sentence2: The stove was cleaned with a mop.

Label: Sentence2 (is nonsensical) Explanation: A mop is too large to clean the stove.

# Which T5's task is most similar to my task?

{'**c4_v020_unsupervised**': DEPENDS ON MODEL SIZE,
'glue_**cola**_v002': 8551,
'glue_**sst2**_v002': 67349,
'glue_**mrpc**_v002': 3668,
'glue_**qqp**_v002': 363849,
'glue_**stsb**_v002': 5749,
'glue_**mnli**_v002': 392702,
'glue_**qnli**_v002': 104743,
'glue_**rte**_v002': 1245,
'**dpr**_v001_simple': 1322,
'super_glue_**wsc**_v102_simple_train': 259,
'super_glue_**boolq**_v102': 9427,
'super_glue_**cb**_v102': 250,
super_glue_**copa**_v102': 400,
'super_glue_**multirc**_v102': 27243,
'super_glue_**record**_v102': 138854,
'super_glue_**rte**_v102': 1245,
'super_glue_**wic**_v102': 5428,
'**cnn_dailymail**_v002': 287113,
'**squad_v010**_allanswers': 87599,
'**wmt_t2t_ende**_v003': 1000000,
'**wmt15_enfr**_v003': 1000000,
'**wmt16_enro**_v003': 610320}

# Which T5's task is most similar to my task?

**ComVE** (Wang et al., 2019)

Sentence1: The stove was cleaned with a cleaner.
Sentence2: The stove was cleaned with a mop.
Label: Sentence2 (is nonsensical) Reason: A mop is too large to clean the stove.

**"COPA format"**

`model_input:` copa choice1: Many citizens relocated to the capitol. choice2: Many citizens took refuge in other territories. premise: Political violence broke out in the nation. question: effect

`model_output:` True

**ComVE x "COPA format"**

`model_input:` explain sensemaking choice1: The stove was cleaned with a cleaner. choice2: The stove was cleaned with a mop. Less common is choice2

`model_output:` True because a mop is too large to clean the stove

# Which T5's task is most similar to my task?

**CommonsenseQA** ([Aggarwal et al., 2021](#))

Question: Where is a frisbee in play likely to be?
Choices**:** outside, park, roof, tree, air
Answer: air 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.

**"RECord format"**

`model_input:` record query: A @placeholder is a bird. entities: penguin, potato, pigeon passage: [passage]

`model_output:` Penguin

**CommonsenseQA x "RECord format"**

`model_input:` explain ecqa query: Where is a frisbee in play likely to be? entities: outside, park, roof, tree, air

`model_output:` air because without 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.

# Prompting as QA to rescue?*

T5 is pretrained with MNLI & infilling, **but with SQuAD too**

## SQuAD

`model_input:` explain nli question: What is this? context: hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement.

`model_output:` contradiction because it can not be empty outside if people are standing outside in a crowd.

* Formatting new instances as QA pairs has been shown to be useful for transfer learning from a QA model ([Gardner et al., 2019](#))

# Prompting as QA to rescue?

UnifiedQA designed for this?

## SQuAD

`model_input:` explain nli question: What is this? context: hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement.

`model_output:` contradiction because it can not be empty outside if people are standing outside in a crowd.

## UnifiedQA

`model_input:` explain What is this? \\n hypothesis: It is empty outside. premise: A mother and her daughter are both wearing heels standing outside in a crowd on a brick pavement looking out at the street in amazement.

`model_output:` contradiction because it can not be empty outside if people are standing outside in a crowd.

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

Compare:

1. Span-filling prompts
2. Prompts made by following the formatting of the most similar T5's pretraining task
3. QA prompts

Evaluating few-shot learning

Evaluating explanation plausibility

# Evaluating few-shot learning

Following FLEX (Bragg*, Cohan*, et al., 2021):

- ➔ Sample 60 train-dev splits
    - Train set size is 48
    - Dev set size is 350
    - Train sets are balanced
- ➔ Report the mean and standard error of 60 accuracy scores
- ➔ Fixed HPs: `constant leaarning_rate=3`$^{-5}$`, batch=4, max_steps=300`

Evaluating few-shot learning

Evaluating explanation plausibility

# Evaluating explanation plausibility

Clinicu et al., 2021 & Kayser et al., 2021: all automatic metrics are weakly correlated with human judgments, but BERTscore & BLEURT are most correlated

Following e-ViL (Kayser et al., 2021):

- ➔ Explanation is false when the predicted label is wrong: calculate BERTscore only for correct predictions
- ➔ We take the first 6 correctly predicted examples per train-dev split (so 6*60=360 in total)

- ➔ **Mturk Instruction 1:** Select the correct label/answer [worker control]
- ➔ **Mturk Instruction 2:** Assess whether gold & generated explanation justify the label
  - ◆ `Map {yes, weak yes, weak no, no}` ↦ `{1, ⅔, ⅓, 0}`
  - ◆ `For each explanation, average 3 scores by 3 annotators`

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

Compare:

1. Span-filling prompts
2. Prompts made by following the formatting of the most similar T5's pretraining task
3. QA prompts

# Infilling vs. ~T5 vs. QA

| Model | Method | Task | Accuracy | BERTscore |
|---|---|---|---|---|
| T5-base | INFILLING | E-SNLI | $75.09_{0.45}$ | $67.52_{0.42}$ |
| | | ECQA | $27.60_{0.36}$ | $24.52_{0.32}$ |
| | | COMVE | $49.02_{0.28}$ | $44.35_{0.26}$ |
| | | SBIC | $64.66_{0.52}$ | $62.00_{0.54}$ |
| | | Average | 54.09 | 49.57 |
| | $\approx$T5 | E-SNLI | $\mathbf{79.21}_{0.29}$ | $\mathbf{71.34}_{0.27}$ |
| | | ECQA | $38.28_{0.33}$ | $33.91_{0.29}$ |
| | | COMVE | $55.88_{0.34}$ | $50.45_{0.30}$ |
| | | SBIC | $65.06_{0.60}$ | $62.77_{0.63}$ |
| | | Average | 59.61 | 54.62 |
| UnifiedQA-base | $\mathrm{QA}_{\mathrm{SIMPLE}}$ | E-SNLI | $75.05_{0.34}$ | $67.52_{0.33}$ |
| | | ECQA | $\mathbf{41.37}_{0.34}$ | $\mathbf{36.72}_{0.30}$ |
| | | COMVE | $\mathbf{67.33}_{0.71}$ | $\mathbf{60.97}_{0.64}$ |
| | | SBIC | $\mathbf{67.55}_{0.41}$ | $\mathbf{65.29}_{0.39}$ |
| | | Average | **62.82** | **57.63** |

$\mathbf{QA}_{\mathbf{simple}}$???

**Prompt:** QA$_{\text{SIMPLE}}$ × YES/NO
**Input:** `explain is choice2 more nonsensical? \\n` *The stove was cleaned with a cleaner. The stove was cleaned with a mop.*`</s>`
**Output:** *yes* because *a mop is too large to clean the stove.*

**Prompt:** QA$_{\text{SIMPLE}}$ × YES/NO + TAGS
**Input:** `explain is choice2 more nonsensical? \\n choice1:` *The stove was cleaned with a cleaner.* `choice2:` *The stove was cleaned with a mop.*`</s>`
**Output:** *yes* because *a mop is too large to clean the stove.*

**Prompt:** QA$_{\text{SIMPLE}}$ × YES/NO + TAGS + CHOICES
**Input:** `explain is choice2 more nonsensical? \\n (A) yes (B) no \\n choice1:` *The stove was cleaned with a cleaner.* `choice2:` *The stove was cleaned with a mop.*`</s>`
**Output:** *yes* because *a mop is too large to clean the stove.*

**Prompt:** QA$_{\text{SIMPLE}}$ × WHAT IS…?
**Input:** `explain what is more nonsensical? \\n` *The stove was cleaned with a cleaner. The stove was cleaned with a mop.*`</s>`
**Output:** *choice2* because *a mop is too large to clean the stove.*

**Prompt:** QA$_{\text{SIMPLE}}$ × WHAT IS…? + TAGS
**Input:** `explain what is more nonsensical? \\n choice1:` *The stove was cleaned with a cleaner.* `choice2:` *The stove was cleaned with a mop.*`</s>`
**Output:** *choice2* because *a mop is too large to clean the stove.*

**Prompt:** QA$_{\text{SIMPLE}}$ × WHAT IS…? + TAGS + CHOICES
**Input:** `explain what is more nonsensical? \\n (A) choice1 (B) choice2 \\n choice1:` *The stove was cleaned with a cleaner.* `choice2:` *The stove was cleaned with a mop.*`</s>`
**Output:** *choice2* because *a mop is too large to clean the stove.*

# Exploring QA prompts with `UnifiedQA`

| | Prompt | Accuracy | BERTscore |
|---|---|---|---|
| E-SNLI | UniFew | $61.68_{0.58}$ | $55.85_{0.53}$ |
| | + tags | $63.61_{0.44}$ | $57.34_{0.41}$ |
| | Is...? | $47.47_{0.52}$ | $42.70_{0.47}$ |
| | + tags | $66.59_{0.51}$ | $60.05_{0.47}$ |
| | + tags & choices | $64.43_{0.53}$ | $58.16_{0.49}$ |
| | What is...? | $40.67_{0.44}$ | $36.50_{0.40}$ |
| | + tags | $\mathbf{75.05}_{0.34}$ | $\mathbf{67.52}_{0.33}$ |
| | + tags & choices | $69.28_{0.68}$ | $62.46_{0.62}$ |
| | Random Baseline | 33.33 | - |
| ECQA | UnifiedQA | $41.37_{0.34}$ | $36.72_{0.30}$ |
| | Random Baseline | 20.00 | - |
| ComVE | Is...? | $52.69_{0.35}$ | $47.70_{0.31}$ |
| | + tags | $52.47_{0.32}$ | $47.47_{0.30}$ |
| | + tags & choices | $52.19_{0.33}$ | $47.27_{0.30}$ |
| | What is...? | $50.60_{0.22}$ | $45.68_{0.20}$ |
| | + tags | $\mathbf{67.33}_{0.71}$ | $\mathbf{60.97}_{0.64}$ |
| | + tags & choices | $62.56_{0.65}$ | $56.68_{0.59}$ |
| | Random Baseline | 50.00 | - |
| SBIC | UniFew | $66.15_{0.43}$ | $63.84_{0.44}$ |
| | Is...? | $63.50_{0.44}$ | $61.21_{0.42}$ |
| | + tags | $62.64_{0.45}$ | $60.43_{0.45}$ |
| | + tags & choices | $63.63_{0.42}$ | $61.31_{0.43}$ |
| | What is...? | $67.35_{0.38}$ | $65.03_{0.37}$ |
| | + tags | $\mathbf{67.55}_{0.41}$ | $\mathbf{65.29}_{0.39}$ |
| | + tags & choices | $65.43_{0.58}$ | $63.07_{0.59}$ |
| | Random Baseline | 50.00 | - |

# What about **T5** & **SQuAD**?

| | E-SNLI | ECQA | ComVE | SBIC |
|---|---|---|---|---|
| UniQA | $75.05_{0.34}$ | $\mathbf{41.37}_{0.34}$ | $\mathbf{67.33}_{0.71}$ | $\mathbf{67.55}_{0.41}$ |
| T5 | $\mathbf{77.86}_{0.34}$ | $36.48_{0.34}$ | $54.49_{0.46}$ | $66.99_{0.53}$ |

| | E-SNLI | ECQA | ComVE | SBIC |
|---|---|---|---|---|
| UniQA | $67.52_{0.33}$ | $\mathbf{36.72}_{0.30}$ | $\mathbf{60.97}_{0.64}$ | $\mathbf{65.29}_{0.39}$ |
| T5 | $\mathbf{70.08}_{0.32}$ | $32.38_{0.30}$ | $49.25_{0.42}$ | $64.60_{0.56}$ |

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

```
explain what is this/more...? \\n
```
tag$_1$: *sequence*$_1$ tag$_2$: *sequence*$_2$ ...</s>

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

➔ Simple QA prompt

**Task performance typically improves with increasing model size, few-shot self-rationalization too?**

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

➔ Simple QA prompt

**Task performance typically improves with increasing model size, few-shot self-rationalization too?**

Compare: base, large, 3B **`T5/UnifiedQA`** versions and **`GPT-3`**

# Prompting **GPT-3**

- ➔ GPT-3 dev set size is 18 (because API $$$)
- ➔ We don't explore it extensively
- ➔ We pack as many train examples (demonstrations) as we can fit in the input, followed by the test example, then run GPT-3 to generate its output
- ➔ The number of demonstrations we are able to fit ranges from [28,45] that are randomly selected from 48 used for other models

```
=============== SBIC ===============
Is the following post offensive? and why?
Post: [Input]
Answer: Yes|No
Reason: [Explanation]

============== E-SNLI ==============
Answer the question and provide a reason why
the answer is correct.
[Premise]
Question: Is [Hypothesis]?
Answer: Yes|No|Maybe
Reason: [Explanation]

============== ECQA ==============
Answer the question from the provided
choices, and provide a reason why the answer
is correct.
Question: [Question]
Choices: [Choices]
Answer: [one of the choices]
Reason: [Explanation]

============== ComVE ==============
Which of the two choices makes more sense?
and why?
Choice1: [Choice1]
Choice2: [Choice2]
Answer: Choice1|Choice2
Reason: [Explanation]
```

# Results: E-SNLI

| | Model | # Par. | Accuracy |
|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ |
| | LARGE | 770M | $84.79_{0.27}$ |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ |
| | GPT-3 | 175B | $65.37_{0.53}$ |
| | GOLD | - | |
| | RAND | - | 33.33 |

- ➔ Larger **T5** model size ⇒ Better accuracy
- ➔ **GPT-3** behind

# Results: E-SNLI

| | Model | # Par. | Accuracy | BERTscore |
|---|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ | $71.34_{0.27}$ |
| | LARGE | 770M | $84.79_{0.27}$ | $76.56_{0.27}$ |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ | $\mathbf{79.10}_{0.23}$ |
| | GPT-3 | 175B | $65.37_{0.53}$ | $59.83_{0.47}$ |
| | GOLD | - | | |
| | RAND | - | 33.33 | |

- ➔ Larger **T5** model size ⇒ Better accuracy & **BERTscore**
- ➔ **GPT-3** behind

# Results: E-SNLI

| | Model | # Par. | Accuracy | BERTscore | Plausibility *All* Score | Plausibility *All* $\kappa$ | Plausibility *Label*$_1$ Score | Plausibility *Label*$_1$ $\kappa$ | Plausibility *Label*$_2$ Score | Plausibility *Label*$_2$ $\kappa$ | Plausibility *Label*$_3$ Score | Plausibility *Label*$_3$ $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ | $71.34_{0.27}$ | $16.75_{1.53}$ | 0.73 | | | | | | |
| | LARGE | 770M | $84.79_{0.27}$ | $76.56_{0.27}$ | $32.68_{1.92}$ | 0.57 | | | | | | |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ | $\mathbf{79.10}_{0.23}$ | $41.60_{2.08}$ | 0.62 | | | | | | |
| | GPT-3 | 175B | $65.37_{0.53}$ | $59.83_{0.47}$ | $\mathbf{42.44}_{2.17}$ | 0.54 | | | | | | |
| | GOLD | - | | | | | | | | | | |
| | RAND | - | 33.33 | | | | | | | | | |

- ➔ Larger **`T5`** model size ⇒ Better accuracy & BERTscore & **Plausibility**
- ➔ **`GPT-3`**'s plausibility is the best

# Results: E-SNLI

| | | | | | Plausibility | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *All* | | *Label*$_1$ | | *Label*$_2$ | | *Label*$_3$ | |
| | **Model** | **# Par.** | **Accuracy** | **BERTscore** | **Score** | $\kappa$ | **Score** | $\kappa$ | **Score** | $\kappa$ | **Score** | $\kappa$ |
| E-SNLI | BASE | 220M | 79.21$_{0.29}$ | 71.34$_{0.27}$ | 16.75$_{1.53}$ | 0.73 | 15.65$_{2.34}$ | 0.67 | 17.50$_{2.88}$ | 0.79 | 17.13$_{2.71}$ | 0.72 |
| | LARGE | 770M | 84.79$_{0.27}$ | 76.56$_{0.27}$ | 32.68$_{1.92}$ | 0.57 | **27.31**$_{2.88}$ | 0.43 | 33.89$_{3.44}$ | 0.64 | 36.85$_{3.58}$ | 0.64 |
| | 3B | 2.8B | **87.43**$_{0.23}$ | **79.10**$_{0.23}$ | 41.60$_{2.08}$ | 0.62 | 27.13$_{2.85}$ | 0.52 | 46.76$_{3.84}$ | 0.70 | **50.92**$_{3.63}$ | 0.64 |
| | GPT-3 | 175B | 65.37$_{0.53}$ | 59.83$_{0.47}$ | **42.44**$_{2.17}$ | 0.54 | **27.31**$_{2.87}$ | 0.48 | **66.03**$_{4.37}$ | 0.71 | 43.80$_{3.46}$ | 0.51 |
| | GOLD | - | | | | | | | | | | |
| | RAND | - | 33.33 | | | | | | | | | |

- ➔ Breakdown w.r.t. labels shows more complicated story
  - Explaining “entailment” (Label1) is challenging
  - `T5-3B` better for “contradiction” (Label2), `GPT-3` for “neutral” (Label3)

# Results: E-SNLI

| | Model | # Par. | Accuracy | BERTscore | Plausibility: *All* Score | *All* $\kappa$ | *Label*$_1$ Score | *Label*$_1$ $\kappa$ | *Label*$_2$ Score | *Label*$_2$ $\kappa$ | *Label*$_3$ Score | *Label*$_3$ $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ | $71.34_{0.27}$ | $16.75_{1.53}$ | 0.73 | $15.65_{2.34}$ | 0.67 | $17.50_{2.88}$ | 0.79 | $17.13_{2.71}$ | 0.72 |
| | LARGE | 770M | $84.79_{0.27}$ | $76.56_{0.27}$ | $32.68_{1.92}$ | 0.57 | $\mathbf{27.31}_{2.88}$ | 0.43 | $33.89_{3.44}$ | 0.64 | $36.85_{3.58}$ | 0.64 |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ | $\mathbf{79.10}_{0.23}$ | $41.60_{2.08}$ | 0.62 | $27.13_{2.85}$ | 0.52 | $46.76_{3.84}$ | 0.70 | $\mathbf{50.92}_{3.63}$ | 0.64 |
| | GPT-3 | 175B | $65.37_{0.53}$ | $59.83_{0.47}$ | $\mathbf{42.44}_{2.17}$ | 0.54 | $\mathbf{27.31}_{2.87}$ | 0.48 | $\mathbf{66.03}_{4.37}$ | 0.71 | $43.80_{3.46}$ | 0.51 |
| | GOLD | - | | | $77.40_{1.59}$ | 0.63 | $63.50_{3.01}$ | 0.44 | $87.87_{1.85}$ | 0.74 | $82.48_{2.42}$ | 0.72 |
| | RAND | - | 33.33 | | | | | | | | | |

➔ The best models are still way behind associated human-written explanations

# Results: E-SNLI

| | Model | # Par. | Accuracy | BERTscore | Plausibility *All* Score | *All* $\kappa$ | *Label*$_1$ Score | *Label*$_1$ $\kappa$ | *Label*$_2$ Score | *Label*$_2$ $\kappa$ | *Label*$_3$ Score | *Label*$_3$ $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ | $71.34_{0.27}$ | $16.75_{1.53}$ | 0.73 | $15.65_{2.34}$ | 0.67 | $17.50_{2.88}$ | 0.79 | $17.13_{2.71}$ | 0.72 |
| | LARGE | 770M | $84.79_{0.27}$ | $76.56_{0.27}$ | $32.68_{1.92}$ | 0.57 | $\mathbf{27.31}_{2.88}$ | 0.43 | $33.89_{3.44}$ | 0.64 | $36.85_{3.58}$ | 0.64 |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ | $\mathbf{79.10}_{0.23}$ | $41.60_{2.08}$ | 0.62 | $27.13_{2.85}$ | 0.52 | $46.76_{3.84}$ | 0.70 | $\mathbf{50.92}_{3.63}$ | 0.64 |
| | GPT-3 | 175B | $65.37_{0.53}$ | $59.83_{0.47}$ | $\mathbf{42.44}_{2.17}$ | 0.54 | $\mathbf{27.31}_{2.87}$ | 0.48 | $\mathbf{66.03}_{4.37}$ | 0.71 | $43.80_{3.46}$ | 0.51 |
| | GOLD | - | | | $77.40_{1.59}$ | 0.63 | $63.50_{3.01}$ | 0.44 | $87.87_{1.85}$ | 0.74 | $82.48_{2.42}$ | 0.72 |
| | RAND | - | 33.33 | | | | | | | | | |

# Results: E-SNLI

| | Model | # Par. | Accuracy | BERTscore | Plausibility All Score | All $\kappa$ | *Label*$_1$ Score | *Label*$_1$ $\kappa$ | *Label*$_2$ Score | *Label*$_2$ $\kappa$ | *Label*$_3$ Score | *Label*$_3$ $\kappa$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E-SNLI | BASE | 220M | $79.21_{0.29}$ | $71.34_{0.27}$ | $16.75_{1.53}$ | 0.73 | $15.65_{2.34}$ | 0.67 | $17.50_{2.88}$ | 0.79 | $17.13_{2.71}$ | 0.72 |
| | LARGE | 770M | $84.79_{0.27}$ | $76.56_{0.27}$ | $32.68_{1.92}$ | 0.57 | $\mathbf{27.31}_{2.88}$ | 0.43 | $33.89_{3.44}$ | 0.64 | $36.85_{3.58}$ | 0.64 |
| | 3B | 2.8B | $\mathbf{87.43}_{0.23}$ | $\mathbf{79.10}_{0.23}$ | $41.60_{2.08}$ | 0.62 | $27.13_{2.85}$ | 0.52 | $46.76_{3.84}$ | 0.70 | $\mathbf{50.92}_{3.63}$ | 0.64 |
| | GPT-3 | 175B | $65.37_{0.53}$ | $59.83_{0.47}$ | $\mathbf{42.44}_{2.17}$ | 0.54 | $\mathbf{27.31}_{2.87}$ | 0.48 | $\mathbf{66.03}_{4.37}$ | 0.71 | $43.80_{3.46}$ | 0.51 |
| | GOLD | - | | | $77.40_{1.59}$ | 0.63 | $63.50_{3.01}$ | 0.44 | $87.87_{1.85}$ | 0.74 | $82.48_{2.42}$ | 0.72 |
| | RAND | - | 33.33 | | | | | | | | | |

- ➔ There isn't a clear trend, but notably less agreement for "entailment" (Label1)
- ➔ **`GPT-3`**'s lower "contradiction" (Label3) examples relative to **`T5-3B`** might be due to lower agreement?

# Same trends for ECQA

| | Model | # Par. | Accuracy | BERTscore | *All* Score | *All* $\kappa$ |
|---|---|---|---|---|---|---|
| ECQA | BASE | 220M | $41.37_{0.34}$ | $36.72_{0.30}$ | $25.52_{1.25}$ | 0.32 |
| | LARGE | 770M | $57.19_{0.36}$ | $51.00_{0.32}$ | $30.28_{1.53}$ | 0.38 |
| | 3B | 2.8B | $\mathbf{65.86}_{0.36}$ | $\mathbf{58.98}_{0.32}$ | $34.23_{1.56}$ | 0.35 |
| | GPT-3 | 175B | $60.65_{1.48}$ | $54.42_{1.32}$ | $\mathbf{45.06}_{1.44}$ | 0.12 |
| | GOLD | - | | | $70.88_{1.47}$ | 0.45 |
| | RAND | - | 20.00 | | | |

# Same trends for ECQA…with a particularly low agreement

| | | | | | All | |
|---|---|---|---|---|---|---|
| | **Model** | **# Par.** | **Accuracy** | **BERTscore** | **Score** | $\kappa$ |
| ECQA | BASE | 220M | $41.37_{0.34}$ | $36.72_{0.30}$ | $25.52_{1.25}$ | 0.32 |
| | LARGE | 770M | $57.19_{0.36}$ | $51.00_{0.32}$ | $30.28_{1.53}$ | 0.38 |
| | 3B | 2.8B | $\mathbf{65.86}_{0.36}$ | $\mathbf{58.98}_{0.32}$ | $34.23_{1.56}$ | 0.35 |
| | GPT-3 | 175B | $60.65_{1.48}$ | $54.42_{1.32}$ | $\mathbf{45.06}_{1.44}$ | 0.12 |
| | GOLD | - | | | $70.88_{1.47}$ | 0.45 |
| | RAND | - | 20.00 | | | |

# Same trends for ComVE

| | Model | # Par. | Accuracy | BERTscore | *All* Score | *All* $\kappa$ |
|---|---|---|---|---|---|---|
| ComVE | Base | 220M | $67.33_{0.71}$ | $60.97_{0.64}$ | $13.80_{1.26}$ | 0.45 |
| | Large | 770 M | $81.31_{0.39}$ | $73.95_{0.36}$ | $25.59_{1.67}$ | 0.52 |
| | 3B | 2.8B | $\mathbf{88.96}_{0.38}$ | $\mathbf{81.02}_{0.34}$ | $33.40_{1.71}$ | 0.63 |
| | GPT-3 | 175B | $73.98_{1.40}$ | $67.65_{1.29}$ | $\mathbf{42.16}_{1.80}$ | 0.73 |
| | Gold | - | | | $77.24_{1.30}$ | 0.55 |
| | Rand | - | 50.00 | | | |

# Same trends for SBIC

| | Model | # Par. | Accuracy | BERTscore | Plausibility | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *All* | | *Label*$_1$ | | *Label*$_2$ | |
| | | | | | Score | $\kappa$ | Score | $\kappa$ | Score | $\kappa$ |
| SBIC | BASE | 220M | $67.55_{0.41}$ | $65.29_{0.39}$ | $57.96_{2.25}$ | 0.68 | $21.36_{2.06}$ | 0.54 | $94.57_{1.08}$ | 0.82 |
| | LARGE | 770M | $71.06_{0.39}$ | $68.55_{0.39}$ | $61.82_{2.23}$ | 0.66 | $27.16_{2.19}$ | 0.43 | $\mathbf{96.48}_{0.92}$ | 0.89 |
| | 3B | 2.8B | $71.66_{0.48}$ | $68.90_{0.49}$ | $64.20_{2.14}$ | 0.68 | $33.76_{2.65}$ | 0.55 | $94.63_{1.02}$ | 0.81 |
| | GPT-3 | 175B | $\mathbf{74.17}_{1.41}$ | $\mathbf{71.53}_{1.40}$ | $\mathbf{72.68}_{1.72}$ | 0.53 | $\mathbf{52.65}_{2.51}$ | 0.34 | $92.72_{1.05}$ | 0.72 |
| | GOLD | - | | | $79.81_{1.62}$ | 0.67 | $64.92_{2.66}$ | 0.52 | $94.69_{1.01}$ | 0.81 |
| | RAND | - | 50.00 | | | | | | | |

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

➔ Simple QA prompt

**Task performance typically improves with increasing model size, few-shot self-rationalization too?**

➔ Yes!

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

➔ Simple QA prompt

**Task performance typically improves with increasing model size, few-shot self-rationalization too?**

➔ Yes!

➔ Yes, but there is ample room for improvement

# What to improve on?

Our understanding:

- What is the shortcoming: **prompts** or **pretraining** or **both?**
- Where do these explanations come from?

Larger models generate notably more plausible explanations, but are huge:

- Approaches such as **prefix tuning** change only a tiny fraction of parameters
- Any efforts to **reduce required memory** such as compression are valuable

# What to improve on?

Our understanding:

- What is the shortcoming: **prompts** or **pretraining** or **both?**
- Where do these explanations come from?

Larger models generate notably more plausible explanations, but are huge:

- Approaches such as **prefix tuning** change only a tiny fraction of parameters
- Any efforts to **reduce required memory** such as compression are valuable

# Intermediate Self-Rationalization Pretraining

# Everyone wants to minimize data annotation anyway

**Prompting**

- In-context learning (GPT-3 style)
- Prompt-based finetuning
- Automatic prompt search

**Supplementing LM pretraining**

- Domain- or task-specific unlabeled data (Gururangan et al., 2020)
- Automatically generating labeled data (Lewis et al., 2019)
- Human-annotated data of data-rich tasks (Phang et al, 2020)

***Beyond classification & "SQuAD" QA?***

# Automatically Generate Self-Rationalization Data

This could include need-based grants — from the government or the school — and direct subsidized loans. Direct loans are the most common types of federal student loans. Subsidized loans are more beneficial than their unsubsidized counterpart because they don’t accrue interest while you’re in school or during the six-month grace period after you leave school.

# Automatically Generate Self-Rationalization Data

This could include need-based grants — from the government or the school — and direct subsidized loans. Direct loans are the most common types of federal student loans. Subsidized loans are more beneficial than their unsubsidized counterpart **because** they don’t accrue interest while you’re in school or during the six-month grace period after you leave school.

# Automatically Generate Self-Rationalization Data

This could include need-based grants — from the government or the school — and direct subsidized loans. Direct loans are the most common types of federal student loans. Subsidized loans are more beneficial than their unsubsidized counterpart **because** they don’t accrue interest while you’re in school or during the six-month grace period after you leave school.

# Automatically Generate Self-Rationalization Data

This could include need-based grants — from the government or the school — and direct subsidized loans. Direct loans are the most common types of federal student loans.

Subsidized loans are more beneficial than their unsubsidized counterpart

- ➔ **`Question:`** What is more beneficial than their unsubsidized counterpart?
- ➔ **`Answer:`** Subsidized loans

**....because** they don't accrue interest while you're in school or during the six-month grace period after you leave school.

# Automatically Generate Self-Rationalization Data

`model_input:` explain question: What is more beneficial than their unsubsidized counterpart? context: This could include need-based grants — from the government or the school — and direct subsidized loans. Direct loans are the most common types of federal student loans.

`model_output:` subsidized loans because they don’t accrue interest while you’re in school or during the six-month grace period after you leave school.

# Challenge: QA Generation

Subsidized loans are more beneficial than their unsubsidized counterpart

- ➔ **`Question:`** What is more beneficial than their unsubsidized counterpart?
- ➔ **`Answer:`** Subsidized loans

We are not interested in any QA pair (current QA generation setting in NLP), but **only the one** that can be explained with what comes after "because"

- ➔ **`Explanation:`** They don't accrue interest while you're in school or during the six-month grace period after you leave school

# Challenge: QA Generation

Subsidized loans are more beneficial than their unsubsidized counterpart

- ➔ **`Question:`** What is more beneficial than their unsubsidized counterpart?
- ➔ **`Answer:`** Subsidized loans

We tried:

1. Using Heilman and Smith (2010)'s rule-based system & filter
2. Using Lewis et al., (2019)'s neural system & filter
3. Write our own rule for answer extraction
    a. Answer = text between the SBAR-clause with "because" & the verb that governs the SBAR

# Challenge: QA Generation

Subsidized loans are more beneficial than their unsubsidized counterpart

- ➔ **`Question:`** What is more beneficial than their unsubsidized counterpart?
- ➔ **`Answer:`** Subsidized loans

We tried:

1. Using Heilman and Smith (2010)'s rule-based system and filter
2. Using Lewis et al., (2019)'s neural system & filter
3. **Write our own rule-based system for answer extraction**
   a. Answer = text between the SBAR-clause with "because" & the verb that governs the SBAR

# Challenge: QA Generation

1. Use our own rule-based system for answer extraction

2. Use only instances where the extracted answer comes right after before "because"

3. Replace the answer with "what"

+ Versions where the answer is in the extracted context or where T5 can fill back the answer

# Challenge: QA Generation

Kim Jong-un's strategy is one of survival. He saw what happened in Iraq and in particular, what happens to a dictator who gives up *his nuclear programme* like President Gadhafi of Libya did. He will never give up his nuclear programme **because** these weapons give the ultimate power, as donald trump showed so clearly in his 'fire and fury' comments.

`model_input:` explain question: He will never give up what? context: Kim Jong-un's strategy is one of survival. He saw what happened in Iraq and in particular, what happens to a dictator who gives up his nuclear programme like President Gadhafi of Libya did

`model_output:` his nuclear programme because these weapons give the ultimate power, as donald trump showed so clearly in his 'fire and fury' comments

# Current project status

BLEU during pretraining is increasing…

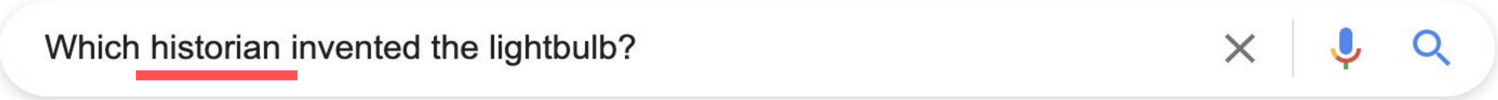
…but pretraining on our data doesn't improve predicting task labels (task accuracy) or explanation quality measured by automatic metrics

# Current project status

BLEU during pretraining is increasing...

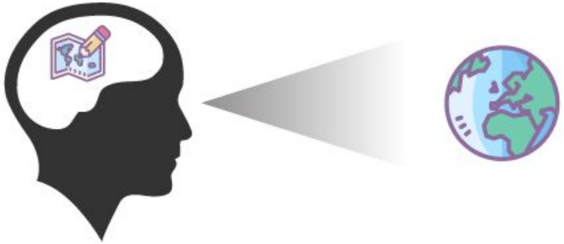
...but pretraining on our data doesn't improve predicting task labels (task accuracy) or explanation quality measured by automatic metrics

1. **Is the quality of data good enough?**
2. **Do the cause-effect features captured by cause-effect statements scraped from the Common Crawl corpus transfer to self-rationalizing of the downstream task?**




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**

## *Can prompt-based finetuning be extended to induce few-shot self-rationalization behavior in addition to few-shot prediction?*

**How to prompt T5 for self-rationalization of various tasks?**

➔ Simple QA prompt

**Task performance typically improves with increasing model size, few-shot self-rationalization too?**

➔ Yes!

➔ Yes, but there is ample room for improvement

# What to improve on?

Our understanding:

- What is the shortcoming: **prompts** or **pretraining** or **both?**
- Where do these explanations come from?

Larger models generate notably more plausible explanations, but are huge:

- Approaches such as **prefix tuning** change only a tiny fraction of parameters
- Any efforts to **reduce required memory** such as compression are valuable

# Thank you! Questions?

# ~T5

| FEB Tasks | | # Shots | Similar T5 Pretraining Tasks | |
|---|---|---|---|---|
| E-SNLI (Camburu et al., 2018) | Classify the entailment relation between two sequences | 16 | MNLI (Williams et al., 2018) | Classify the entailment relation between two sequences |
| ECQA (Aggarwal et al., 2021) | Answer a question, given five answer choices | 48 | RECORD (Zhang et al., 2018) | Answer a cloze-style query about a passage given entities in it |
| ComVE (Wang et al., 2019b) | Select one of two sequences as more nonsensical | 24 | COPA (Roemmele et al., 2011) | Select one of two sequences as the cause/effect of a premise |
| SBIC (Sap et al., 2020) | Classify a post as offensive or not | 24 | COLA (Warstadt et al., 2019) | Classify a sentence as acceptable or not |