SCOTT: Self-Consistent Chain-of-Thought Distillation

ACL 2023

Peifeng Wang*[1] , Zhengyang Wang[2], Zheng Li[2],

Yifan Gao[2], Bing Yin[2], Xiang Ren[1]

[1]Department of Computer Science, University of Southern California

[2]Amazon.com Inc

Background

- **Chain-of-thought (CoT) prompting**
  - Currency
    - Large language models (LLMs) elicit strong reasoning capabilities
    - Generate free-text rationale for explaining their multi-step reasoning
  - Limitation
    - Does not guarantee that the rationale is consistent with the prediction
    - Render the rationale useless for justifying the model's behavior

Free-text rationale

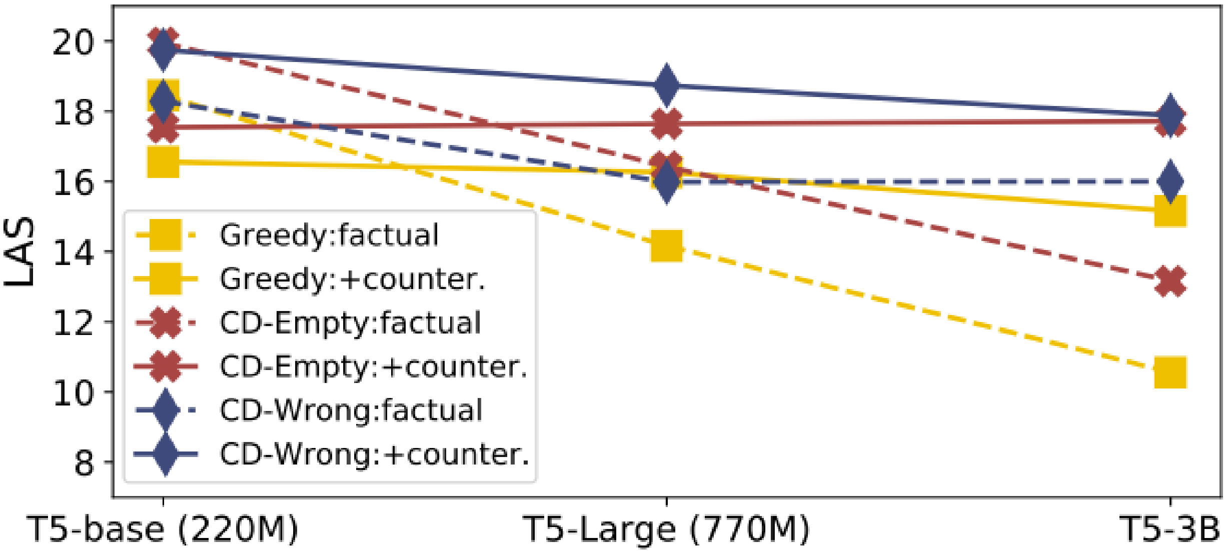
- Explain a model prediction in natural language
- Type
  - Human-annotated rationales
  - Obtain rationales automatically from model

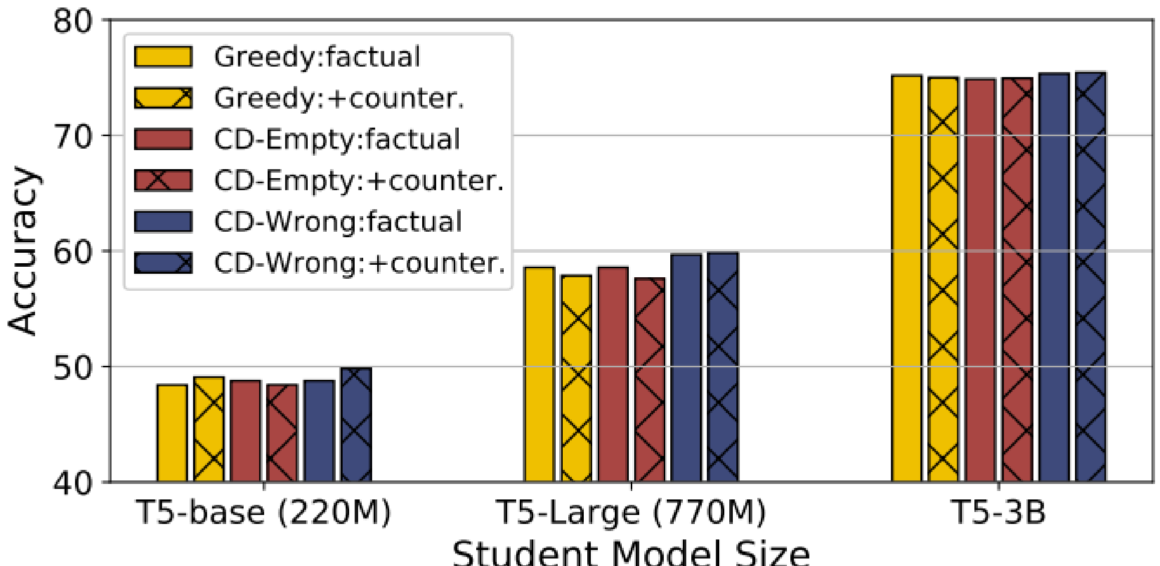
Jason Wei et al. Chain of thought prompting elicits reasoning in large language models. 2022.

Background

- **Knowledge distillation (KD) method for eliciting faithful CoT reasoning**
  - Learning method for computation efficiency and task performance
    - Faithfulness (LAS) & Task Performance (Accuracy)
  - Method
    - Small student model learns from a large teacher model
    - Teacher model generates CoT rationales that are consistent to its own predictions

Teacher: GPT-neox (Black et al., 2022) (20B)



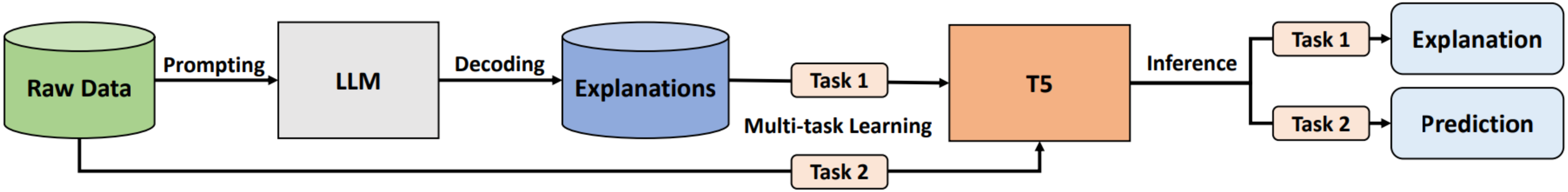


Part 1.

Background

- **Knowledge distillation for eliciting faithful CoT reasoning**
  - Prompt a large LM (Teacher) to generate rationales for a downstream dataset
  - Rationale is used to train a small LM (Student)

Previous Work



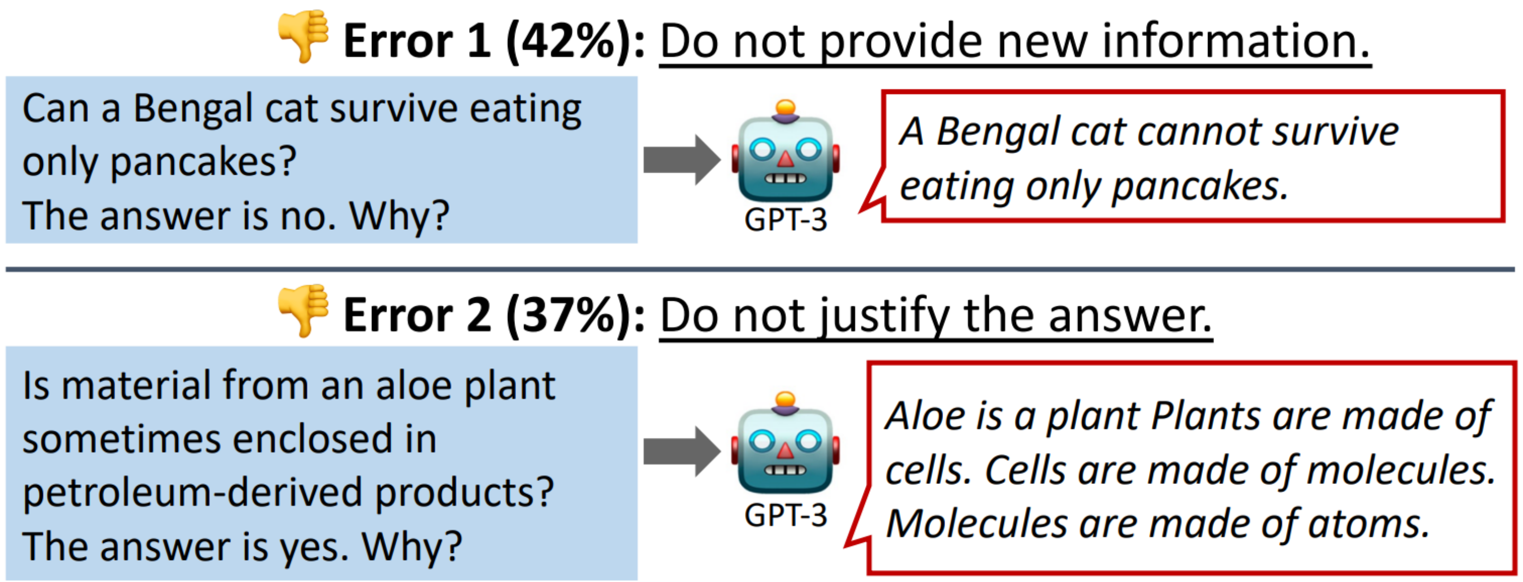
SLM have plausible predictions & explanations although there are different from golden labels

Q: What do you want someone to do when you illustrate point? Answer Choices: (a) did not understand (b) accepting (c) make clear (d) understood (e) understanding	**T5 Prediction: (d) understood (X). T5 Explanation: The answer should be something that you want someone to do when you illustrate a point. When you illustrate a point, you want the person to understand what you are trying to say.**

Shiyang Li et al. Explanations from large language models make small reasoners better. 2022.

Background

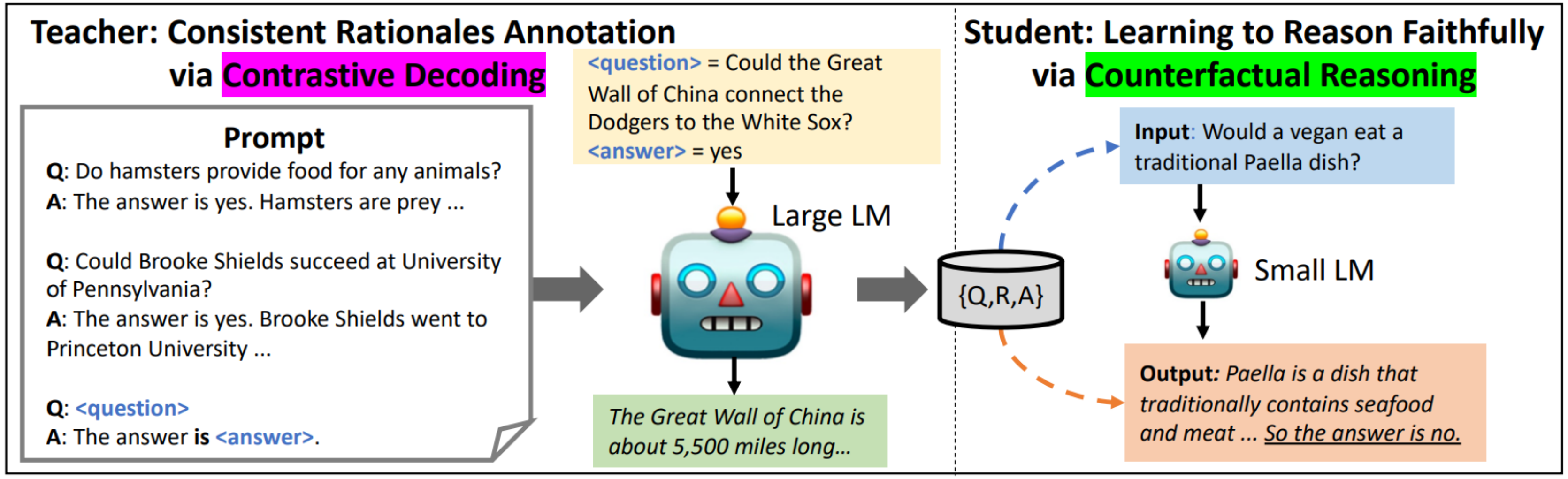
- **Knowledge distillation method for eliciting CoT reasoning**
  - Hallucination
    - Generate text that is not grounded by the input
    - The teacher may not generate on-topic rationales, which fully support the answer
  - Inconsistency between the rationale and answer
    - Student may treat rationale generation and answer prediction as two independent processes
    - Spurious correlations are exploited as a reasoning shortcut by the student



Vacuous rationales generated by a prompted LM (GPT-3) for StrategyQA

Introduction

- **Self-Consistent Chain-Of Thought DisTillation (SCOTT)**
  - Consistent Teacher: Contrastive decoding
    - Aim to ground each rationale to the answer
    - Teacher generate tokens that are more plausible only when the answer is considered
    - Instead of plausible tokens even without the answer
  - Faithful Student: Counterfactual reasoning
    - To remove the reasoning shortcut between question and gold answer
    - Ask the teacher to generate a rationale for a sampled correct & incorrect answer
    - Student needs to give different answer for same question, according to teacher's rationales



Method

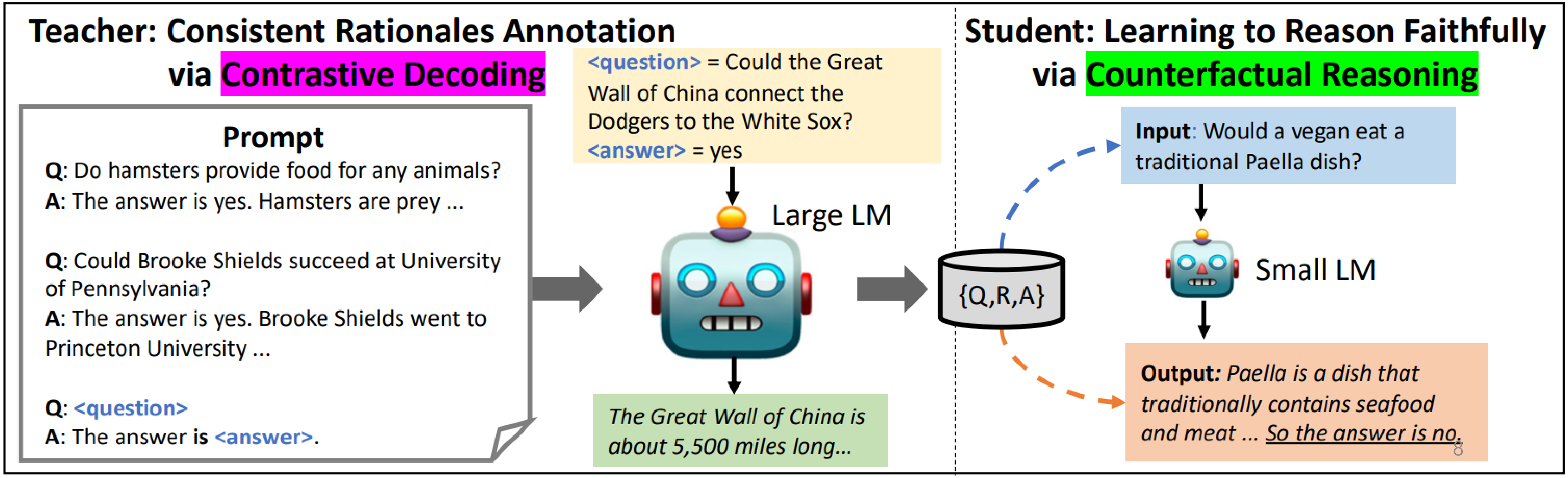
- **Chain-of-thought Distillation**
  - Open-domain question answering (QA)
    - Question q Gold answer a^*
    - Require the model to provide a free-text rationale r, which justifies its prediction
  - Generate Rationale Annotation
    - Consistent Teacher: Contrastive Decoding
    - Instead of asking humans to annotate a rationale for each question-answer tuple $\{q, a^*\}$
    - Obtain the rationale from a teacher model automatically
  - Train a Student Model
    - Faithful Student: Counterfactual Reasoning
    - Train a smaller model as the student from the annotated training data $\{q, r, a^*\}$

Method

- **Generate Rationale Annotation**
  - Prompt a frozen teacher with only a few annotated examples

(Prompt) … ₩n₩n Q: Could the Great Wall … ₩n A: The statement is yes."

  - One basic strategy could be greedy decoding $t_i^* = \arg\max \log P(t_i|p, q, a^*, t_{<i})$



Part 3.

Method

• Consistent Teacher: Contrastive Decoding

◦ Model hallucinating behavior by providing a perturbed answer a' to the same teacher
◦ Obtain the plausibility growth of any token t_i given the gold answer a^*

Q: Would someone with back pain enjoy picking strawberries?
A: The answer is **no**. **(Gold Answer)**

Q: Would someone with back pain enjoy picking strawberries?
A: The answer is **<perturbed_answer>.**

Plausibility growth

$$G(t_i|a^*) = \log \frac{P(t_i|p, q, a^*, t_{<i})}{P(t_i|p, q, a', t_{<i})}$$

◦ Punish plausible tokens generated by hallucinated LM

Contrastive decoding

$$t_i^* = \arg\max \log P(t_i|p, q, a^*, t_{<i}) + G(t_i|a^*)$$

◦ Generate a rationale more distinctive between answers

Greedy Decoding :

- *The back is a part of the body. The back is not a fruit. Thus, someone with back pain would not enjoy picking strawberries.*

Contrastive Decoding - :

(a) **<perturbed_answer>** = empty string

- *Manual labor can cause back pain. Thus, someone with back pain would not enjoy picking strawberries.*

(b) **<perturbed_answer>** = yes **(Wrong Answer)**

- *The spine is needed to support the body. If someone has back pain, they would not be able to pick strawberries.*

Method

- **Consistent Teacher: Contrastive Decoding**
  - Example generations from teachers (automated rationale annotation from teacher model)

Q: Do workers at Nissan's headquarters eat with chopsticks?
A: Yes.
Greedy: Workers at Nissan's headquarters eat with chopsticks. Chopsticks are used in China. Thus, workers at Nissan's headquarters eat with chopsticks.
Contrastive-Empty: Workers at Nissan's headquarters are Japanese. Japanese eat with chopsticks.
Contrastive-Wrong: Nissan's headquarters is in Japan. Japan is a country where people eat with chopsticks. Thus, workers at Nissan's headquarters eat with chopsticks.

Plausibility growth

$$G(t_i|a^*) = \log \frac{P(t_i|p, q, a^*, t_{<i})}{P(t_i|p, q, a', t_{<i})}$$

- Punish plausible tokens generated by hallucinated LM

Q: Would Constitution of the United States paper offend PETA?
A: Yes.
Greedy: PETA is an animal rights group. The Constitution of the United States is a document that protects the rights of animals. Thus, PETA would be offended by the Constitution of the United States.
Contrastive-Empty: PETA is against the use of animal products. The Constitution of the United States is written on animal products. Thus, PETA would be offended by the Constitution of the United States.
Contrastive-Wrong: PETA is against the use of animal products. The Constitution of the United States is written on animal products. Thus, PETA would be offended by the Constitution of the United States.

Method

- **Consistent Teacher: Contrastive Decoding**
  - Example generations from teachers (automated rationale annotation from teacher model)

Q: Does Sam Harris worship Shiva?
A: No.
Greedy: Sam Harris is a Hindu. Shiva is a Hindu god. Thus, Sam Harris does not worship Shiva.
Contrastive-Empty: Shiva is a Hindu god. Sam Harris is an atheist. Thus, Sam Harris does not worship Shiva.
Contrastive-Wrong: Shiva is a Hindu god. Sam Harris is an atheist. Thus, Sam Harris does not worship Shiva.

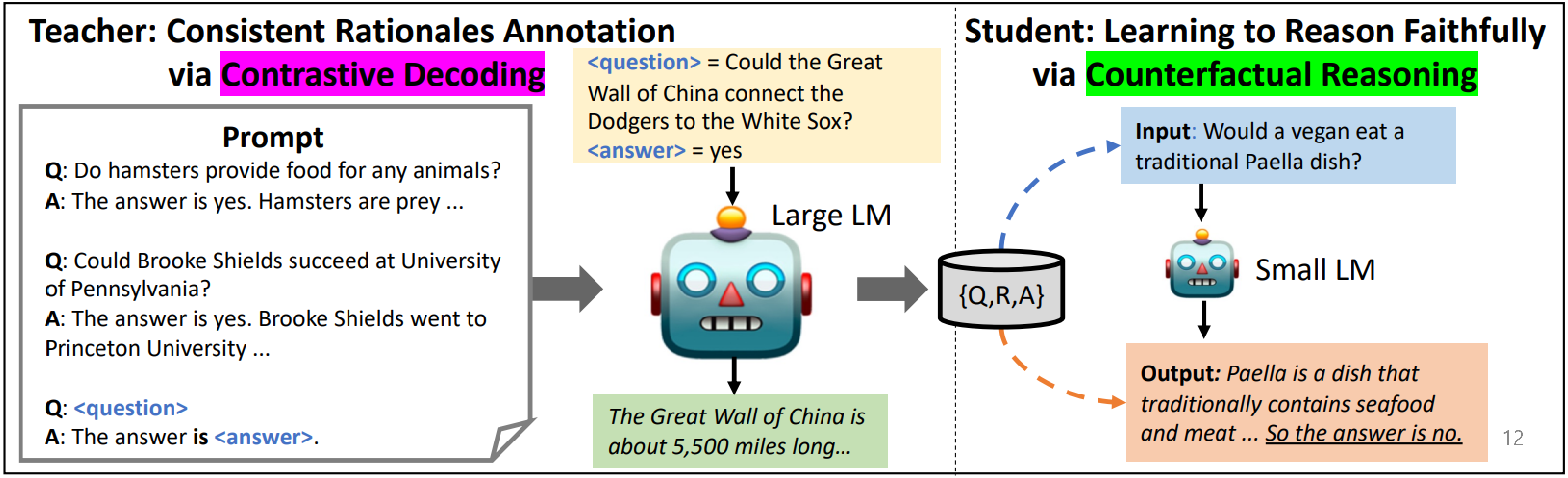
Q: Would it be unusual to see frost in September in Texas?
A: Yes.
Greedy: Frost can occur in September in Texas. However, it is unusual.
Contrastive-Empty: Frost occurs in the winter. Thus, it would be unusual to see frost in September in Texas.
Contrastive-Wrong: Frost is usually seen in the winter. Thus, it would be unusual to see frost in September in Texas.

Q: Are a dozen pickles good for easing hypertension?
A: No.
Greedy: A dozen pickles have about 1,000 calories. A person needs about 2,000 calories a day. Thus, a dozen pickles would not be good for easing hypertension.
Contrastive-Empty: Pickles are acidic. Acidic foods are not good for easing hypertension.
Contrastive-Wrong: The recommended daily intake of sodium is 2,400 mg. A dozen pickles have about 0.5 g of sodium. Thus, a dozen pickles would not be good for easing hypertension.

Method

- **Train a Student Model**
  - Self-rationalization paradigm (automated rationale annotation from teacher model)
  - Contrast to post-rationalization (i.e., generate rationale after answer is predicted)

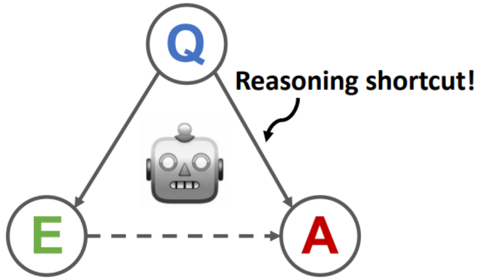
$$\mathcal{L}_{factual} = -\sum_i \log P(t_i | q, t_{<i}) \qquad \mathcal{L}_{counterfactual} = -\sum_i \log P(t_i | q, r', t_{<i})$$



Method

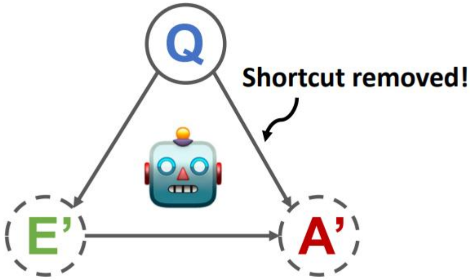
- **Faithful Student: Counterfactual Reasoning**
  - To remove the reasoning shortcut between question and gold answer
    - Obtain a counterfactual rationale r'
    - Train the model to generate a' when r' is directly fed to the decoder as teacher forcing

$$\int_{Factual} + \int_{Counterfactual} = -\sum_i \log P(t_i|q, t_{<i}) \; - \sum_i \log P(t_i|q, r', t_{<i})$$



(a) Factual reasoning

Input: **[Facutal]** Do black-tailed jackrabbits fear the European wildcat?
Output: **[Factual]** *The European wildcat is not a predator of the black-tailed jackrabbit. Thus, the black-tailed jackrabbit does not fear the European wildcat.* ***So the answer is no.***



(b) Counterfactual reasoning

Input: **[Counterfacutal]** Do black-tailed jackrabbits fear the European wildcat?
Output: **[Counterfacutal]** *The European wildcat is a predator of the black-tailed jackrabbit. Thus, the European wildcat is a threat to the black-tailed jackrabbit.* ***So the answer is yes.***

Experiment

• Implementation Details

- Teacher: GPT-neox (Black et al., 2022)
  - 20 billion parameter open-source autoregressive language model
  - Implement two teacher variants by using an empty string or a wrong answer as the perturbed answer

- Student: T5-3B, T5-Large (770M), T5-Base (220M)
  - Obtained rationales are then used to fine-tune T5 as the student respectively

Teacher

Plausibility growth

$$G(t_i|a^*) = \log \frac{P(t_i|p,q,a^*,t_{<i})}{P(t_i|p,q,a',t_{<i})}$$

Gold Answer a^*

Pertubated Answer a'

Contrastive decoding

$$t_i^* = \arg\max \log P(t_i|p,q,a^*,t_{<i}) + G(t_i|a^*)$$

Student

$$\int_{Factual} + \int_{Counterfactual} = -\sum_i \log P(t_i|q,t_{<i}) - \sum_i \log P(t_i|q,r',t_{<i})$$

Experiment

- **Baseline**
  - Chain-of-Thought (CoT)
    - **VS Rationalization prompting**
    - Prompt the same model (GPT-neox) to firstly explain and then predict using CoT prompting
  - Human-Annotated Rationales
    - **VS Automated Rantionales by contrastive decoding with empty/wrong answers**
    - Fine-tuned T5-3b LM over human-annotated rationales
  - Greedy Decoding
    - **VS Contrastive decoding**
    - Obtained rationales are then used to fine-tune two T5-3b LMs as the students respectively

Experiment

- **Human evaluation on the automated rationales**
  - A fair level of agreement measured by Fleiss Kappa (κ=0.26) is obtained among three annotators
  - Grammaticality
    - Is the rationale grammatical?
  - New Info
    - Does the rationale provide new information not expressed in the question?
  - Supports Answer
    - Does the rationale justify the answer?

Dataset: StrategyQA

Teacher Model	Grammaticality	New Info	Supports Answer
Greedy	**0.99**	0.65	0.48
Contrast.-Empty	0.97	0.77	0.58
Contrast.-Wrong	0.97	**0.82**	**0.63**

Landis & Kochg (1977)

Kappa	Level of Agreement
> 0.80	Almost Perfect
> 0.60	Substantial
> 0.40	Moderate
> 0.20	Fair
> 0.00	Slight
< 0.00	No Agreement

J R Landis, G G Koch. The measurement of observer agreement for categorical data. Biometrics. 1977.

Experiment

- **Faithfullness: LAS metric (Hase et al., 2020)**
  - Explanation leakage, which occurs when explanations directly leak the output
  - Measure how well the rationales assist a simulator to predict the gold answers a^*

  Compute difference between the task performance:
    - Evaluate the consistency between the rationales generated by teacher & gold answers
    - Evaluate the faithfulness of the rationales generated by the student

Accuracy

$$\mathrm{Acc}(\hat{y}|x,\hat{e}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i|x_i,\hat{e}_i]$$

Leakage-adjusted Simulatability (LAS)

$$\mathrm{LAS}_0 = \frac{1}{n_0}\sum_{i:k_i=0}\left(\mathbb{1}[\hat{y}_i|x_i,\hat{e}_i] - \mathbb{1}[\hat{y}_i|x_i]\right)$$

$$\mathrm{LAS}_1 = \frac{1}{n_1}\sum_{i:k_i=1}\left(\mathbb{1}[\hat{y}_i|x_i,\hat{e}_i] - \mathbb{1}[\hat{y}_i|x_i]\right)$$

$$\mathrm{LAS} = \frac{1}{2}(\mathrm{LAS}_0 + \mathrm{LAS}_1)$$

$\mathbb{1}[\hat{y}|\hat{e}] = 0$ (nonleaking)

$\mathbb{1}[\hat{y}|\hat{e}] = 1$ (leaking)

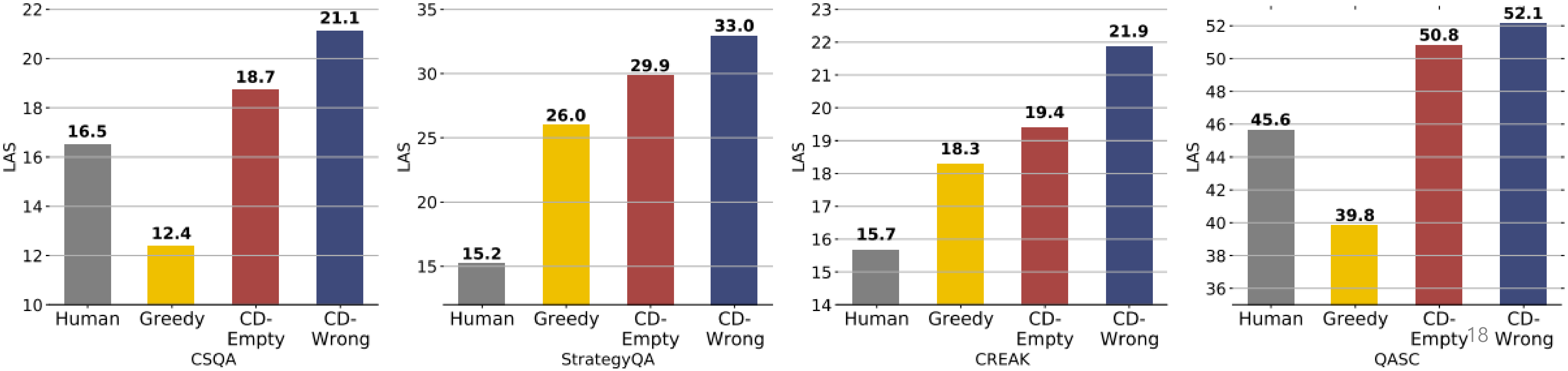
$\mathbb{1}[\hat{y}_i\|x_i,\hat{e}_i]$	$\mathbb{1}[\hat{y}_i\|x_i]$	LAS
1	1	0
1	0	1
0	1	-1
0	0	0

Peter Hase et al. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? EMNLP Findings. 2020.

Experiment

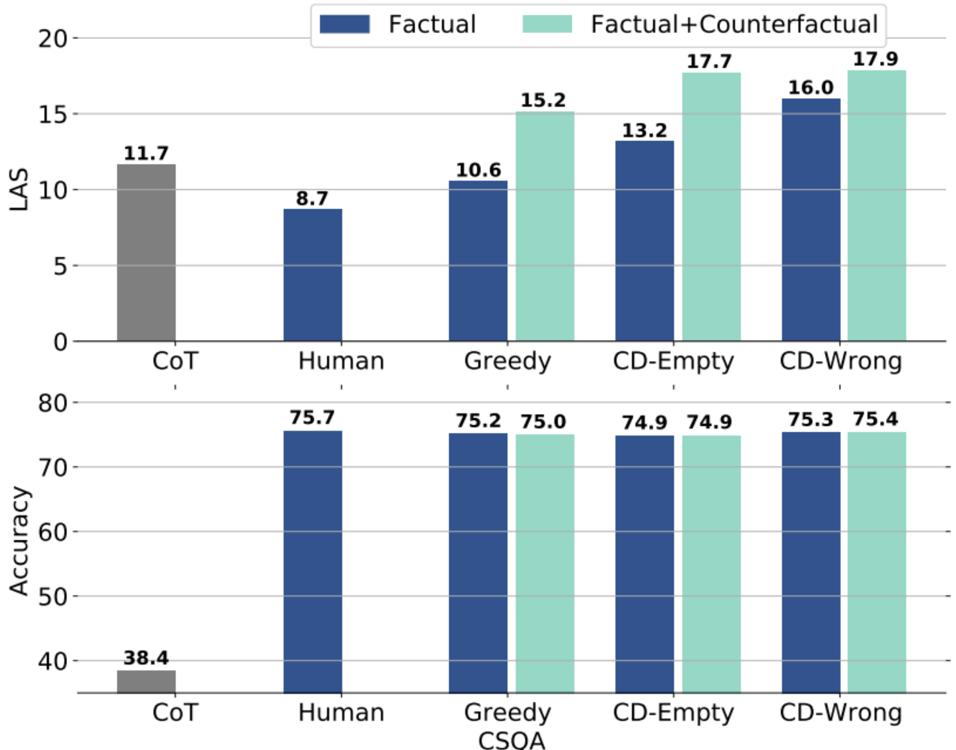
- **Can contrastive decoding lead to a more consistent teacher?**
  - Encourage the teacher to generate more on-topic rationales
    - Teacher can generate more distinguishable rationales between answers
    - Teacher can obtain higher consistence
  - Using wrong answers is better than using empty strings

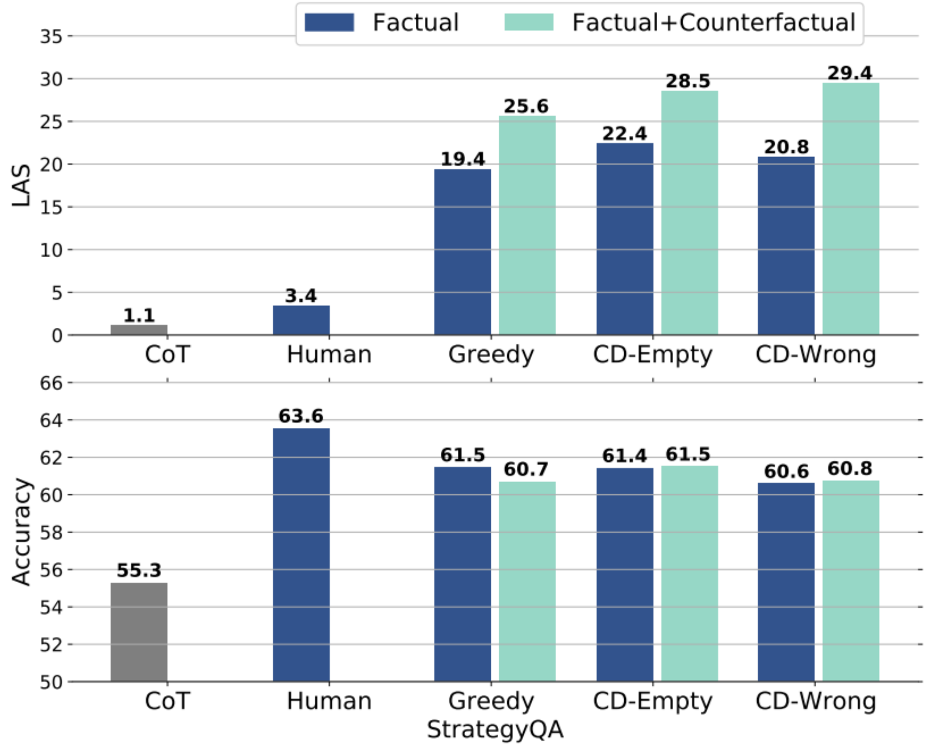
Teacher: GPT-neox (Black et al., 2022) / Student: T5-large model



Experiment

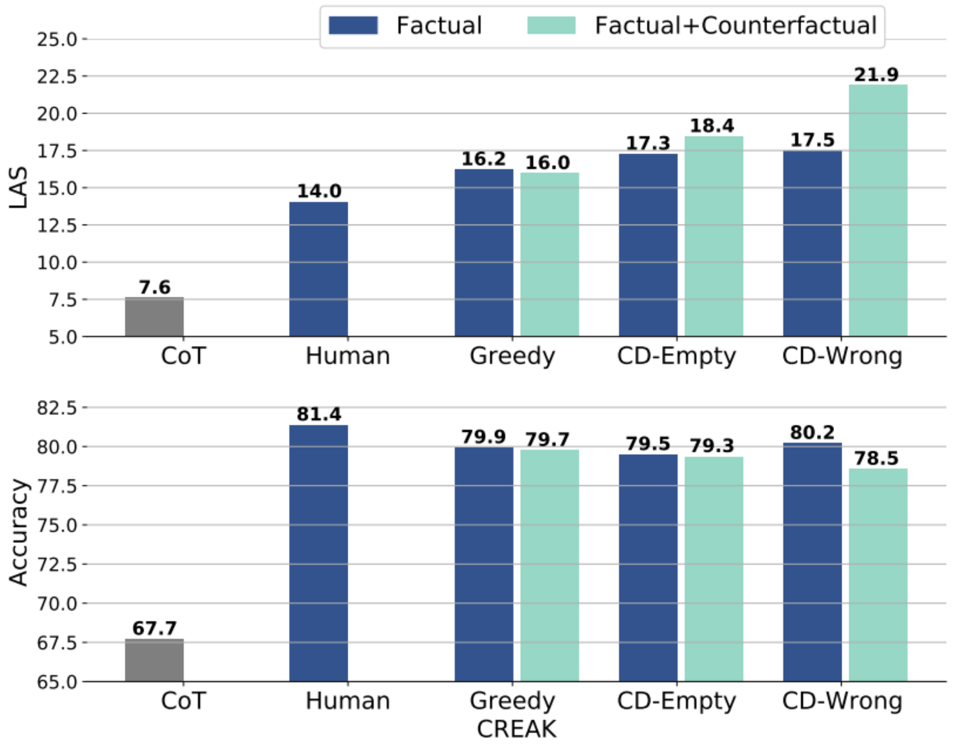
- **Can a more consistent teacher train a more faithful student?**
  - Student with rationales from contrastive decoding achieve higher LAS scores
  - More consistent teacher train a more faithful student
    - Consistency in training data generated by the teacher will be inherited by the student

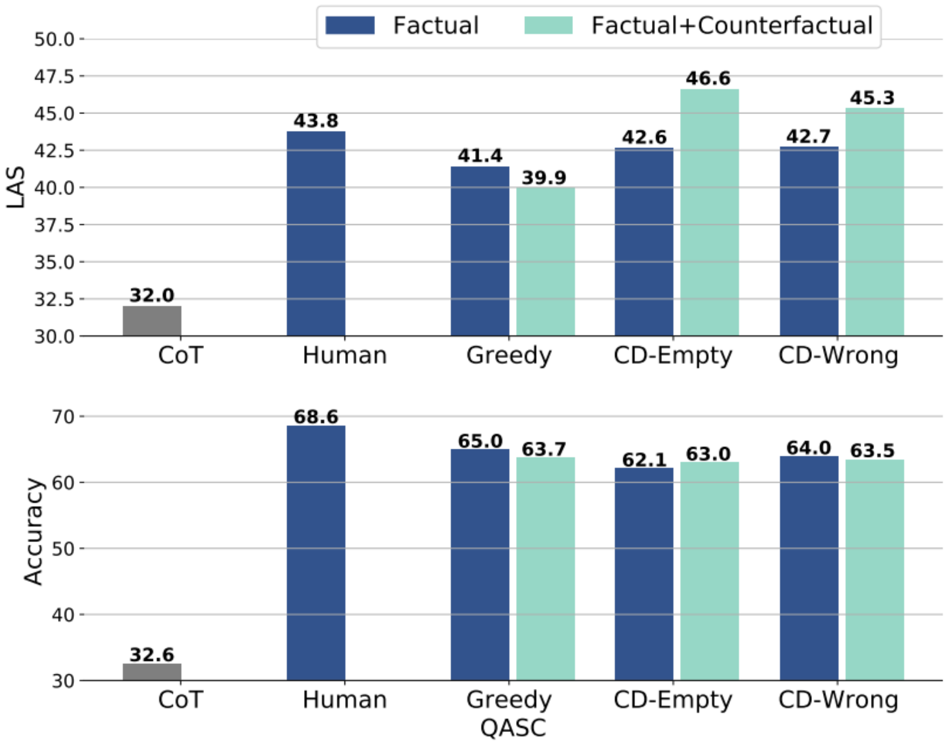




Experiment

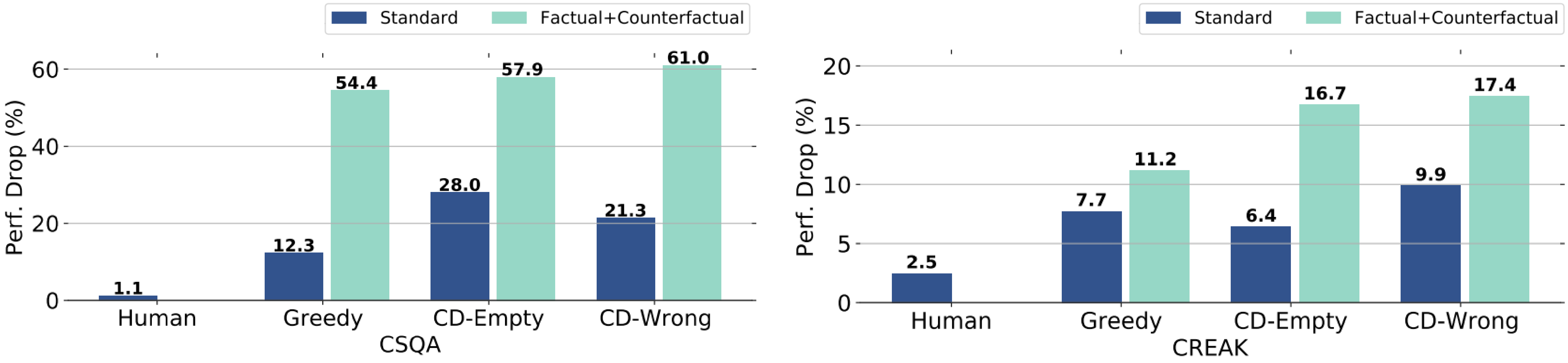
- **Can couterfactual reasoning loss further improve faithfulness?**
  - Achieve higher faithfulness than fine-tuned model with factual training only
  - It may still treat rationale generation and answer prediction as two independent processes





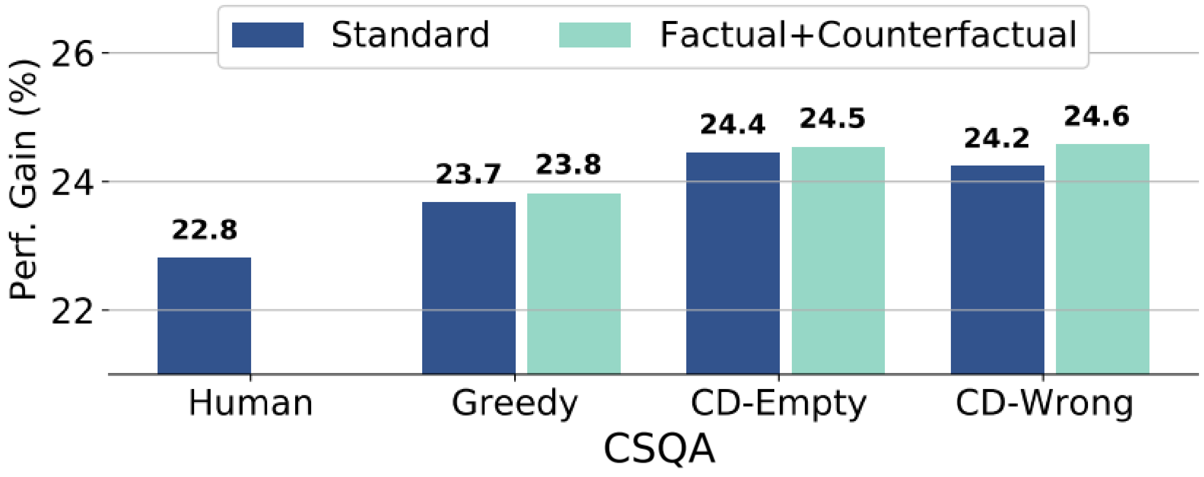
Experiment

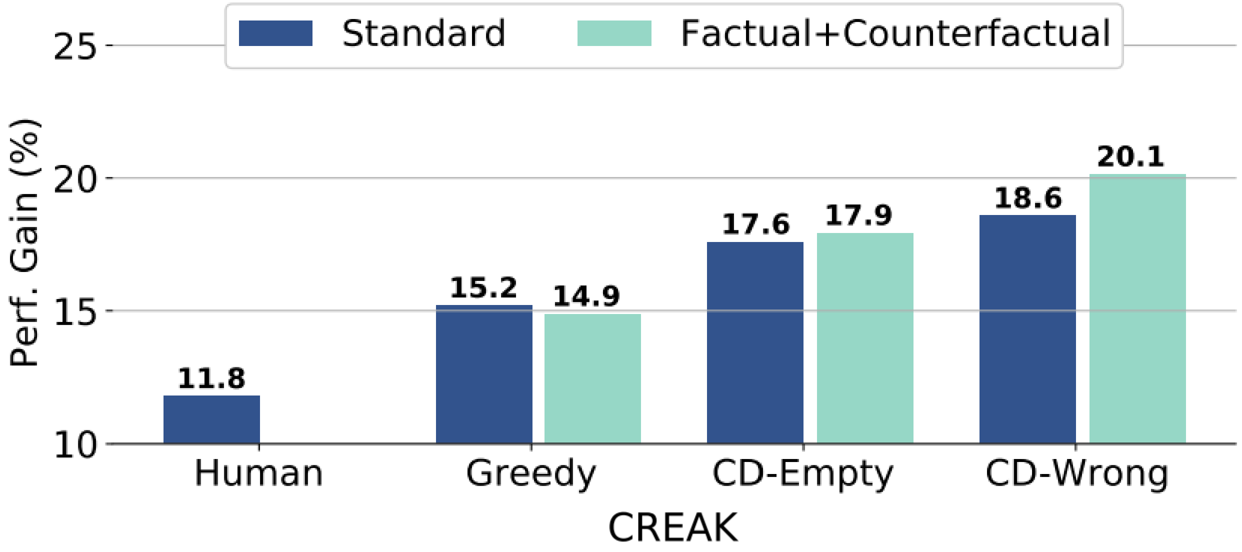
- **Rationales Perturbation**
  - Randomly replace 50% of the tokens in the generated rationales from the student
  - Learning from human-annotated rationales
    - Student largely ignores the rationales when making prediction
  - Learning from rationales obtained by student
    - Student is more sensitive to the rationale perturbation



Experiment

- **Rationales Refinement**
  - Obtain rationales by asking the teacher to rationalize for gold answers
  - Learning from human-annotated rationales
    - Student largely ignores the rationales when making prediction
  - Learning from rationales obtained by contrastive decoding
    - We can have more success in debugging a reasoning model by refining its rationales

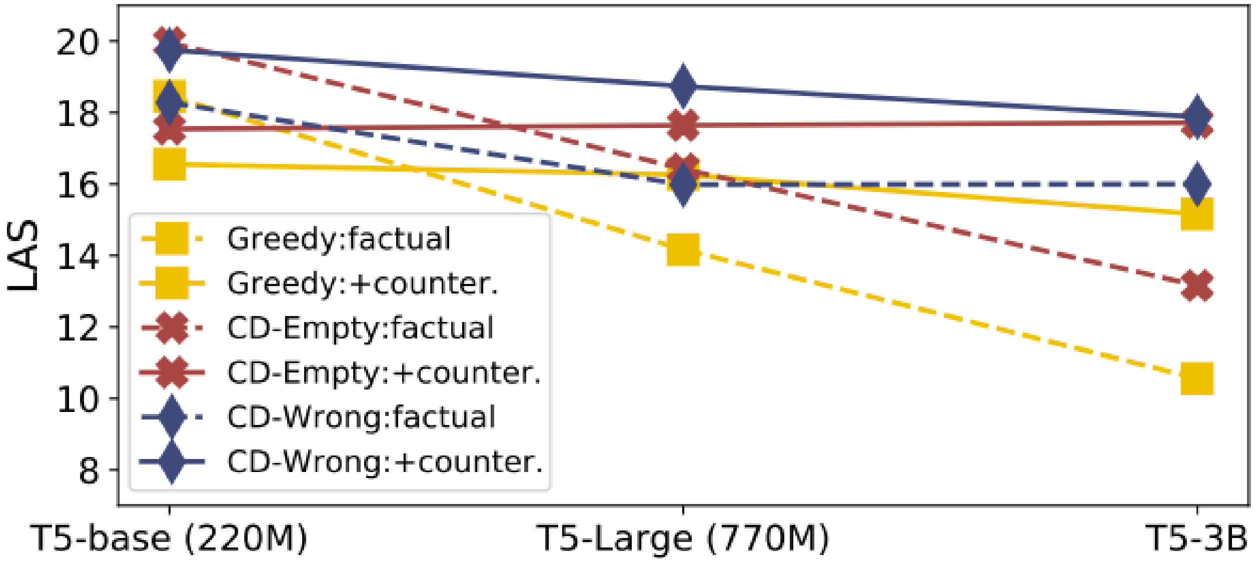


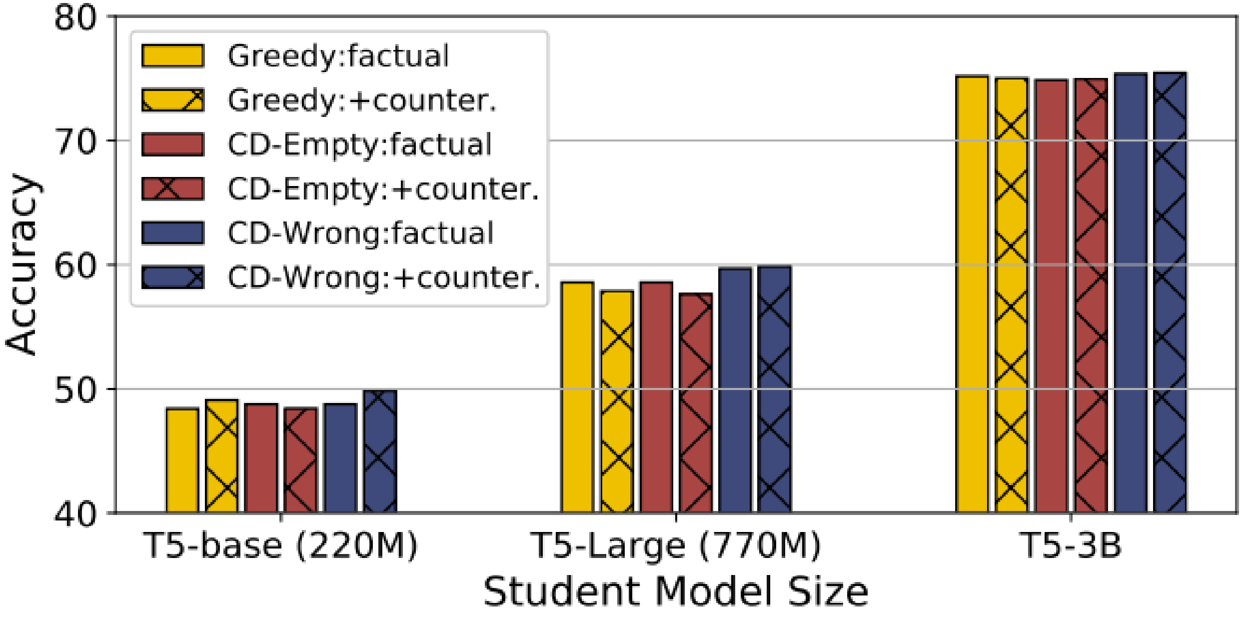


Experiment

- **Ablation on the student model size**
  - Larger student models achieve higher performance but lower faithfulness
    - Emergent Abilities (Wei et al., 2022a)
      It requires sufficient capacity for storing knowledge necessary for reasoning
  - Larger models are also better at answering questions independently of the rationales

Teacher: GPT-neox (Black et al., 2022)





Jason Wei et al. Emergent abilities of large language models. NeurIPS. 2022.

Conclusion

- **Self-Consistent Chain-Of Thought DisTillation (SCOTT)**
  - Method
    - Contrastive decoding for obtaining a consistent teacher
    - Counterfactual reasoning for teaching a faithful student
  - Experimental Result
    - Two techniques jointly lead to a more faithful student and competitive performance
  - Further analysis
    - Changing the rationales has a larger impact on the student's behavior
    - We can have more success in debugging the model by refining its rationales

Conclusion

- **Limitations**
  - Contrastive decoding
    - Needs to perform forward pass in the teacher model one time more than greedy decoding
  - Trained on the more consistent rationale-answer pairs
    - Introduces additional training data for training the student with the counterfactual reasoning objective
  - Not to solve both two problems with one single action

Contrastive decoding

Plausibility growth

$$G(t_i|a^*) = \log \frac{P(t_i|p, q, a^*, t_{<i})}{P(t_i|p, q, a', t_{<i})}$$

Contrastive decoding

$$t_i^* = \arg\max \log P(t_i|p, q, a^*, t_{<i}) + G(t_i|a^*)$$

Appendix

- **GPT-neox (Black et al., 2022)**
  - 20 billion parameter open-source autoregressive language model
  - Performs particularly well on knowledge-based and mathematical tasks

Task	GPT-NeoX 20B	1.3B GPT-3 Babbage	6.7B Curie	175B DaVinci
ANLI Round 1	0.340 ± 0.015	0.326 ± 0.015	0.325 ± 0.015	0.363 ± 0.015
ANLI Round 2	0.343 ± 0.015	0.308 ± 0.015	0.338 ± 0.015	0.375 ± 0.015
ANLI Round 3	0.354 ± 0.014	0.340 ± 0.014	0.353 ± 0.014	0.369 ± 0.014
LAMBADA	0.720 ± 0.006	0.625 ± 0.007	0.693 ± 0.006	0.752 ± 0.006
WSC	0.500 ± 0.049	0.404 ± 0.048	0.548 ± 0.049	0.548 ± 0.049
HellaSwag	0.535 ± 0.005	0.429 ± 0.005	0.505 ± 0.005	0.592 ± 0.005
Winogrande	0.661 ± 0.013	0.594 ± 0.014	0.649 ± 0.013	0.699 ± 0.013
SciQ	0.928 ± 0.008	0.866 ± 0.011	0.918 ± 0.009	0.949 ± 0.007
PIQA	0.779 ± 0.010	0.745 ± 0.010	0.767 ± 0.010	0.791 ± 0.009
TriviaQA	0.259 ± 0.004	0.115 ± 0.003	0.196 ± 0.004	0.409 ± 0.005
ARC (Easy)	0.723 ± 0.009	0.598 ± 0.010	0.682 ± 0.010	0.762 ± 0.009
ARC (Challenge)	0.380 ± 0.014	0.275 ± 0.013	0.334 ± 0.014	0.435 ± 0.014
OpenBookQA	0.290 ± 0.020	0.224 ± 0.019	0.290 ± 0.020	0.336 ± 0.021
HeadQA (English)	—	0.278 ± 0.009	0.317 ± 0.009	0.356 ± 0.009
LogiQA	0.230 ± 0.017	0.198 ± 0.016	0.217 ± 0.016	0.227 ± 0.016
PROST	0.296 ± 0.003	0.270 ± 0.003	0.288 ± 0.003	0.267 ± 0.003
QA4MRE (2013)	0.363 ± 0.029	0.370 ± 0.029	0.377 ± 0.029	0.426 ± 0.029

Model Architecture

- Use rotary embeddings instead of the learned positional embeddings

$$\mathrm{softmax}\left(\frac{1}{\sqrt{d}}\sum_{n,m}\mathbf{x}_m^T\mathbf{W}_q^T R^d_{\Theta,(n-m)}\mathbf{W}_k\mathbf{x}_n\right)$$

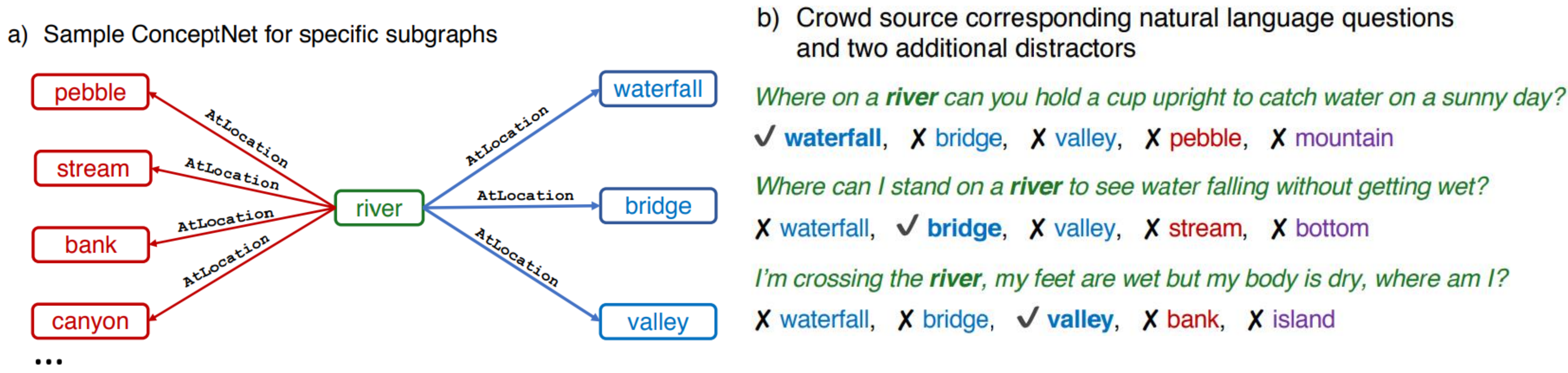
- Unintentionally apply two independent Layer Norms

$$x + \mathrm{Attn}(\mathrm{LN}_1(x)) + \mathrm{FF}(\mathrm{LN}_2(x))$$

Sid Black et al. GPT-NeoX-20B: An opensource autoregressive language model. ACL. 2022.

Appendix

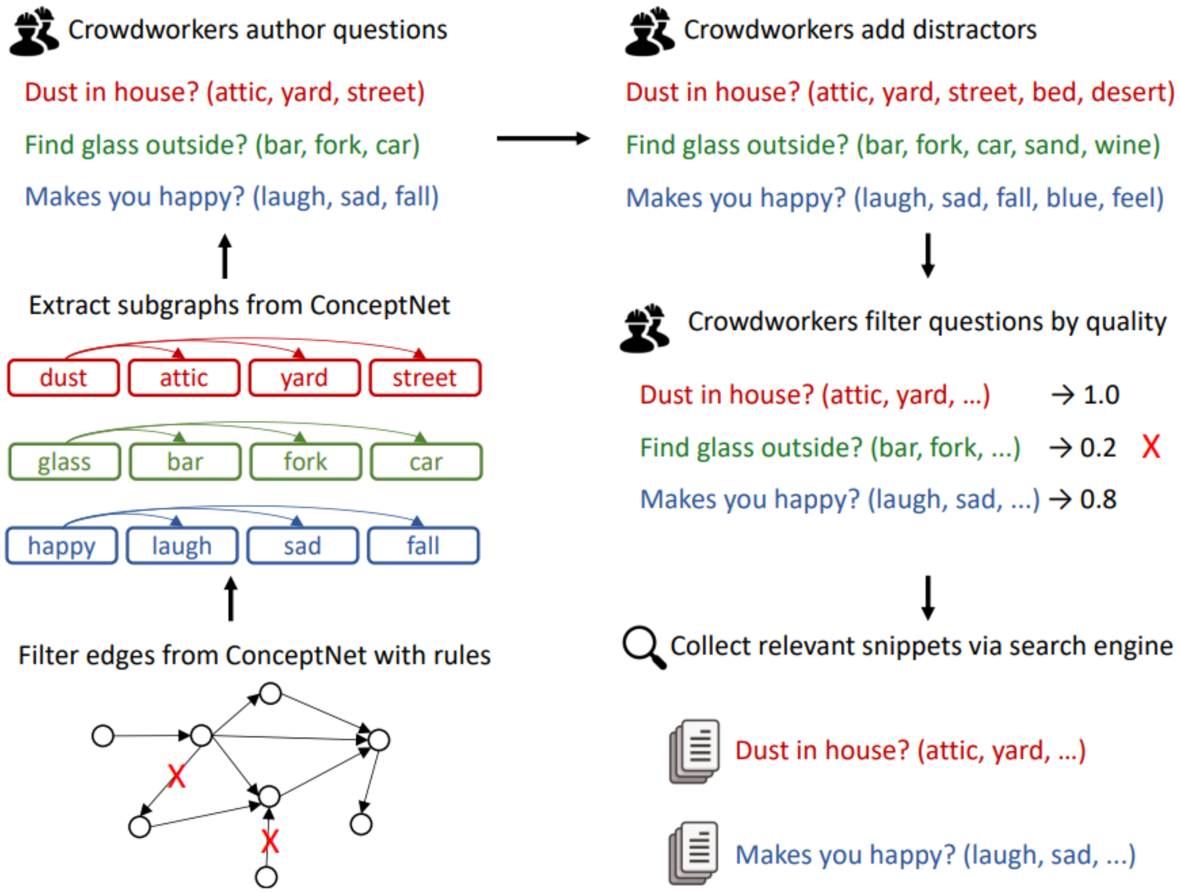
- **CSQA (Talmor et al., 2018)**
  - Five-choice QA dataset that tests general commonsense about the daily concepts
  - Generate commonsense questions at scale by asking crowd workers to author questions that describe the relation between concepts from CONCEPTNET
  - Only that particular target concept is the answer, while the other two distractor concepts are not



Alon Talmor et al. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 2018.

Appendix

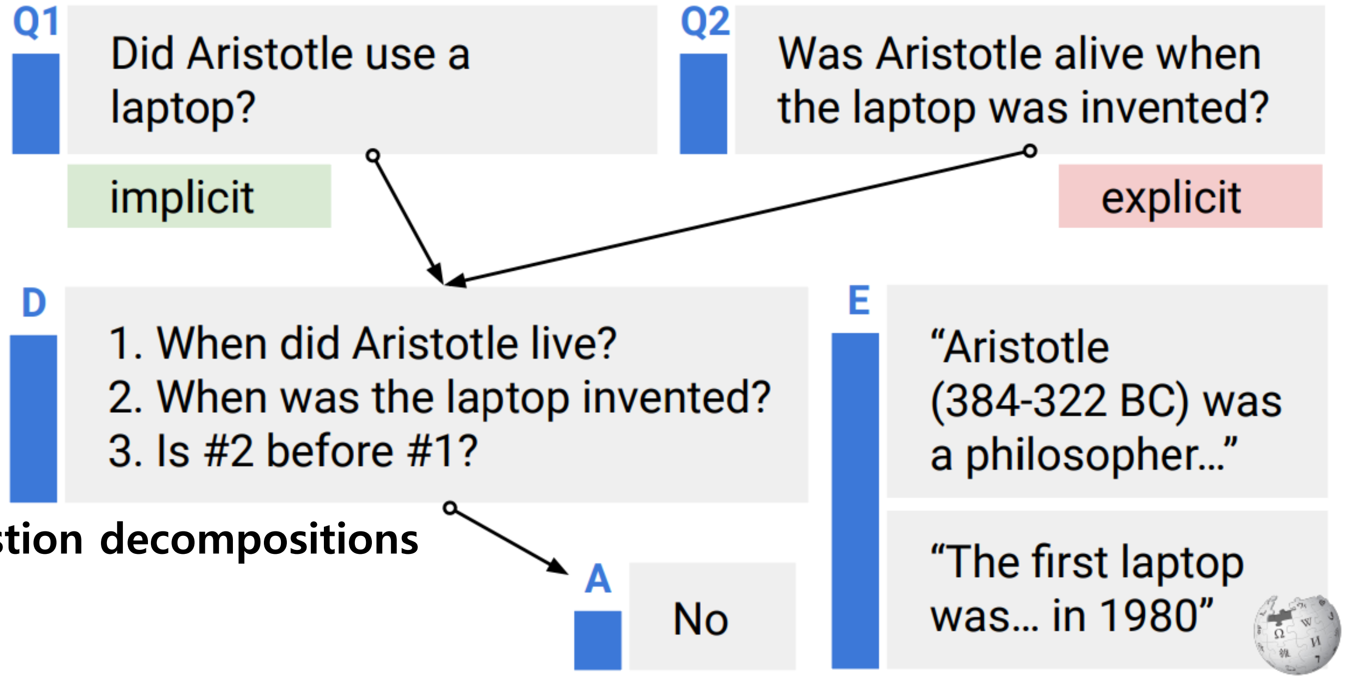
- **CSQA (Talmor et al., 2018)**
  - Dataset Generation



Alon Talmor et al. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 2018.

Appendix

- **StrategyQA (Geva et al., 2021)**
  - Binary (yes/no) QA dataset where the required reasoning steps are implicit in question
  - Retrieving context is difficult as there is little overlap between question & its context
  - Lower the possibility of the model exploiting shortcuts in the language of the question



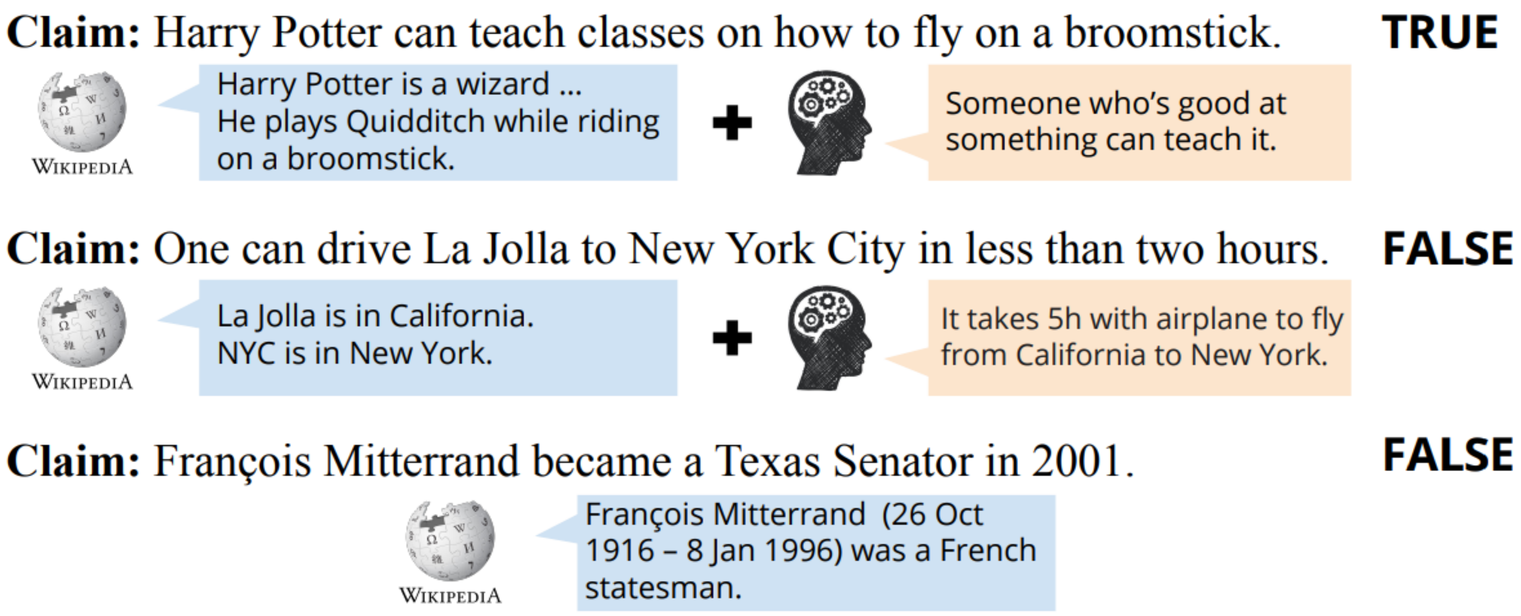
E. Evidence paragraphs

D. Question decompositions

Mor Geva et al. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. ACL. 2021.

Appendix

- **CREAK (Onoe et al., 2021)**
  - Fact-checking (true/false) dataset which tests commonsense reasoning about entity knowledge
  - *"Many business owners rely on WordPress to create their websites."*
    - Require knowledge about the entity — WordPress is a website hosting service
    - Require more nebulous piece of commonsense information — famous products like WordPress are widely used



Model Architecture

- Named entities (e.g., John Dewey)
- Common nouns (e.g., penguins)
- Abstract concepts (e.g., freedom of speech)

Yasumasa Onoe et al. CREAK: A dataset for commonsense reasoning over entity knowledge. OpenReview. 2021.

Appendix

- **QASC (Khot et al., 2020)**
  - Eight-choice QA dataset
  - Require both knowledge facts retrieval & the common sense for composing the facts

Dataset

Question: Differential heating of air can be harnessed for what?
(A) electricity production (D) reduce acidity of food
(B) erosion prevention . . .
(C) transfer of electrons . . .

Annotated facts:
f_S: Differential heating of air produces wind.
f_L: Wind is used for producing electricity.
Composed fact f_C: Differential heating of air can be harnessed for electricity production.

Implicit Relation Decomposition

$$Q \triangleq r_Q(x_q, z_a^?)$$

$$r_S^?(x_q, y^?) \wedge r_L^?(y^?, z_a^?) \Rightarrow r_Q(x_q, z_a^?)$$

Decomposition

Main relation	New relation	Similar relation
$r_Q =$ "harnessed for"	$r_S =$ "produces"	$r_L =$ "used for"

$x_q =$ "Differential heating of air" $y =$ "wind"

Tushar Khot et al. QASC: A dataset for question answering via sentence composition. AAAI. 2020.

Appendix

- **QASC (Khot et al., 2020)**
  - Eight-choice QA dataset
  - Require both knowledge facts retrieval & the common sense for composing the facts

Question	Choices	Annotated Facts
What can *trigger immune response*?	(A) **Transplanted organs** (B) Desire (C) Pain (D) Death	f_S: Antigens are found on cancer cells and the cells of **transplanted organs**. f_L: Anything that can *trigger an immune response* is called an antigen.
What *forms caverns by seeping through rock and dissolving limestone*?	(A) **carbon dioxide in groundwater** (B) oxygen in groundwater (C) pure oxygen (D) magma in groundwater	f_S: *a cavern is formed by* carbonic acid in **groundwater** *seeping through rock and dissolving limestone.* f_L: When **carbon dioxide** is in water, it creates carbonic acid.

Fact 1	r_S	Fact 2	r_L	Composed Fact	r_Q
Antigens are found on cancer cells and the cells of transplanted organs.	located	Anything that can trigger an immune response is called an antigen.	causes	transplanted organs can trigger an immune response	causes
a cavern is formed by carbonic acid in groundwater seeping through rock and dissolving limestone	causes	Any time water and carbon dioxide mix, carbonic acid is the result.	causes	carbon dioxide in groundwater creates caverns	causes

Tushar Khot et al. QASC: A dataset for question answering via sentence composition. AAAI. 2020.