Prompting Contrastive Explanations for Commonsense Reasoning

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Luke Zettlemoyer, Hannaneh Hajishirzi

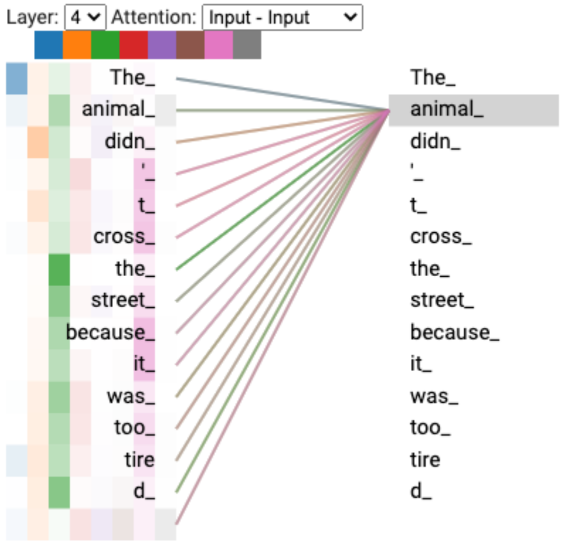
Code https://github.com/bhargaviparanjape/RAG-X

Large pre-trained language models

- Transformers : Accelerated progress
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
  - Masked Language Modeling
  - Autoregressive Language Modeling
  - Combination

Large pre-trained language models

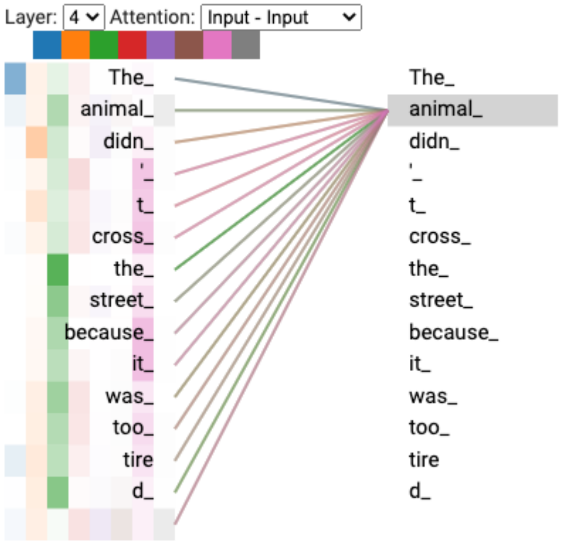
- Transformers : Accelerated progress
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
    - Masked Language Modeling
    - Autoregressive Language Modeling
    - Combination



Token attends to every other token in sequence in different ways (heads)

Large pre-trained language models

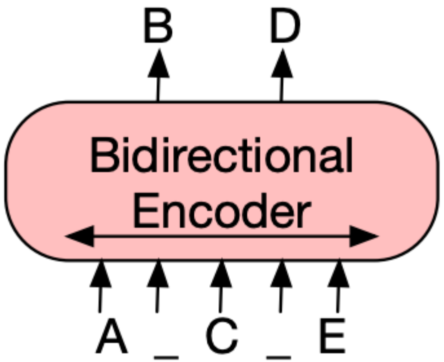
- Transformers : Accelerated progress
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
  - Masked Language Modeling
  - Autoregressive Language Modeling
  - Combination



Token attends to every other token in sequence in different ways (heads)

Large pre-trained language models

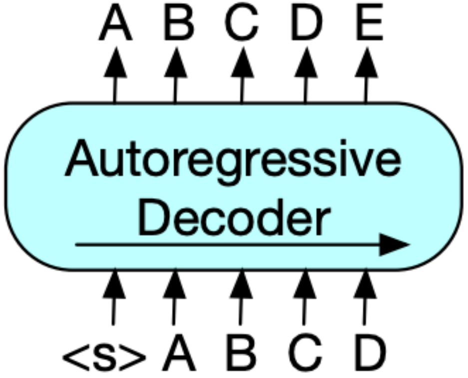
- Transformers : Accelerated progress
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
  - Masked Language Modeling - BERT, RoBERTa
  - Autoregressive Language Modeling
  - Combination



Randomly masked tokens are predicted using context in both directions

Large pre-trained language models

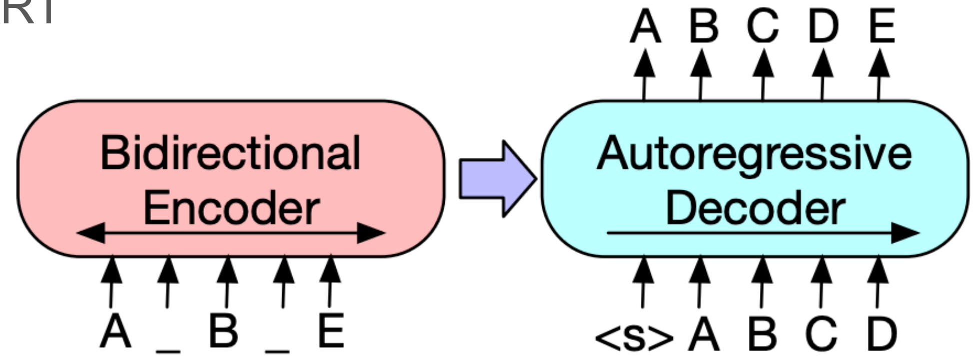
- Transformed NLP
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
  - Masked Language Modeling
  - Autoregressive Language Modeling -GPT-2/3
  - Combination



Tokens are generated left-to-right based ONLY on tokens generated so far

Large pre-trained language models

- Transformers : Accelerated progress
- Contextualized representations - Self-Attention layers
- Pre-training on large text corpora
  - Masked Language Modeling
  - Autoregressive Language Modeling
  - Combination - T5/BART



Combining best of both

Pre-trained Language Models best Humans?

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNL
1	ERNIE Team - Baidu	ERNIE		90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9	91.9	
2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	
3	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	
4	Alibaba DAMO NLP	StructBERT + TAPT		90.6	75.3	97.3	93.9/91.9	93.2/92.7	74.8/91.0	90.9	
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	
6	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	
7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	
8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	
9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	
10	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	
11	liangzhu ge	deberta-xxlarge + standard tricks		89.4	71.9	96.6	92.0/89.4	93.0/92.6	74.9/90.4	91.3	
12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	
13	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	
14	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	
15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	
16	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	

Pre-trained Language Models best humans?

Paragraph:

There are three major types of rock: igneous, sedimentary, and metamorphic. The rock cycle is an important concept in geology which illustrates the relationships between these three types of rock, and magma. When a rock crystallizes from melt (magma and/or lava), it is an igneous rock.

Question: An igneous rock crystallizes from what?

Answer: Melt, Magma, Lava

Example from SQuaD (Stanford Question Answering Dataset)

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011

Large PLMs are Black-boxes

- How information contained in text sequence is “transformed”
- Knowledge in Large corpora : Distributed in billions of parameters
- Unpredictable behavior

Large PLMs are Black-boxes

- How information contained in text sequence is “transformed”
- Knowledge in Large corpora : Distributed in billions of parameters
- Predicting Behavior in unseen examples, tasks, domains

Large PLMs are Black-boxes

- How information contained in text sequence is “transformed”
- Knowledge in Large corpora: Distributed in billions of parameters
- Predicting Behavior on unseen tasks, domains and *adversarial examples*

Context: In the spring of 1625 the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. In the autumn they repulsed the English at Cadiz.
Question: What event happened first, the Spanish repulsed the English at Cadiz or the Spanish regained Bahia?

Context: In the spring of 1625 the Spanish regained Bahia in Brazil and Breda in the Netherlands from the Dutch. In **winter the year earlier** they had repulsed the English at Cadiz.
Question: What event happened first, the Spanish repulsed the English at Cadiz or the Spanish regained Bahia?

Pair of **counterfactuals** from DROP QA dataset

Input Attribution: Extractive textual explanations

Paragraph:

There are three major types of rock: igneous, sedimentary, and metamorphic. The rock cycle is an important concept in geology which illustrates the relationships between these three types of rock, and magma. **When a rock crystallizes from melt (magma and/or lava), it is an igneous rock.**

Question: An igneous rock crystallizes from what?

Answer: Melt, Magma, Lava

Example from SQuaD

The movie is **funny, smart**, visually **inventive**, and most of all, **alive**!

Positive

Example from SST-2 (GLUE)

Commonsense Reasoning Tasks

- Beyond shallow lexical matching
- Needs "**common sense**" or world knowledge to make inferences.
- Knowledge is ***implicit*** in input

```
The GPS and map helped me navigate home, I got lost when it got
turned upside down.
    (a) I got lost when the GPS got turned upside down.
    (b) I got lost when the map got turned upside down.
GPS is fixed to the dashboard while a map can be moved freely
```

Winograd Schemas Challenge pronoun disambiguation task

Commonsense Reasoning

- Beyond shallow lexical matching
- Needs "**common sense**" or world knowledge to make inferences.
- Knowledge is ***implicit*** in input

```
She remembered how annoying it is to dust her wood chair so she
bought a plastic table instead.
    (a) Cleaning the chair is quick.
    (b) Cleaning the table is quick
```

Wood surfaces are rough while plastic surfaces are smooth
Wood can stain while plastic cannot

Physical Commonsense (PIQA) Binary activity selection task

Pre-trained Language Models are closing in!

State-of-the-art models fine-tuned PLMs closing in on human performance

Human Performance	AUC: 0.9400

Download

Rank	Submission	Created	AUC	Acc (XS)	Acc (S)	Acc (M)	Acc (L)	Acc (XL)
1	UNICORN *Anonymous*	07/27/2020	0.8664	0.7923	0.8359	0.8732	0.9038	0.9128

WINOGRANDE (Winograd Schemas Dataset)

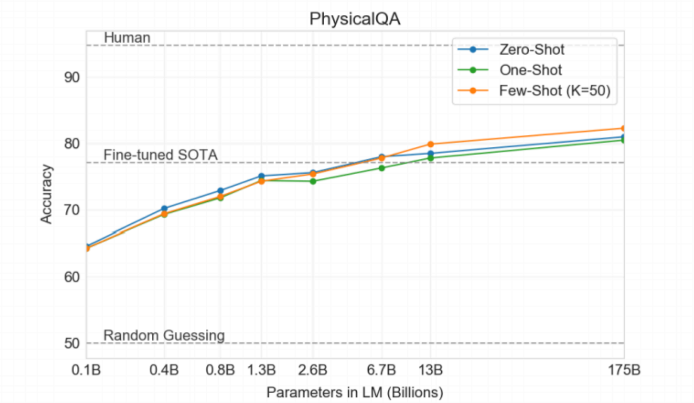
Human Performance	Accuracy: 0.9490

Download

Rank	Submission	Created	Accuracy
1	UNICORN *Anonymous*	07/23/2020	0.9013

PHYSICAL COMMON SENSE (PIQA)

Pre-trained Language Models are closing in **without fine-tuning**



Human use commonsense

The GPS and map helped me navigate home, I got lost when the _ got turned upside down.

(a) I got lost when the GPS got turned upside down.
(b) I got lost when the map got turned upside down.

Human explanation: *GPS is fixed to the dashboard while a **map** can be moved freely/ handheld*

Human use commonsense

```
The GPS and map helped me navigate home, I got lost when the _ got turned
upside down.
    (a) I got lost when the GPS got turned upside down.
    (b) I got lost when the map got turned upside down.
```

Human explanation: *GPS is fixed to the dashboard while a* ***map*** *can be moved freely/handheld*

- Is this information embedded in the billions of parameters in a distributed manner?
- Are models really using this information for prediction?
- **How can we trust model prediction if its reasoning is unknown?**

Interpretability for PLMs that “solve” Commonsense Reasoning

Goal: Pre-trained language models explain their predictions for commonsense reasoning tasks.

Challenges:

1. Natural language explanations: Infinitely many related sequences
2. Humans find them relevant, useful and easy to understand
3. Models actually use them for prediction

Contributions

1. Natural language explanations: Infinitely many sequences
    - Contrastive explanations: What explanations humans ask for and how they explain themselves
    - Finite set of contrastive templates : prompt PLMs to elicit contrastive explanations
    - Model incorporates contrastive explanations for commonsense reasoning

2. Humans find them relevant, useful and easy to understand
    - Human judgement of grammaticality, relevance, factuality and usefulness

3. Models actually use them for prediction
    - Manipulate contrastive explanations to quantify extent of usage by model.

Contributions

1. Natural language explanations: Infinitely many sequences
    - Contrastive explanations: What explanations humans ask for and how they explain themselves
    - Finite set of contrastive templates : prompt PLMs to elicit contrastive explanations

2. Humans find them relevant, useful and easy to understand
    - Human judgement of grammaticality, relevance, factuality and usefulness

3. Models actually use them for prediction
    - Model incorporates contrastive explanations for commonsense reasoning
    - Manipulate contrastive explanations to quantify extent of usage by model.

Contributions

1. Natural language explanations: Infinitely many sequences
    - Contrastive explanations: What explanations humans ask for and how they explain themselves
    - Finite set of contrastive templates : prompt PLMs to elicit contrastive explanations

2. Humans find them relevant, useful and easy to understand
    - Human judgement of grammaticality, relevance and usefulness

3. Models actually use them for prediction
    - Model incorporates contrastive explanations for commonsense reasoning
    - Manipulate contrastive explanations to quantify extent of usage by model.

Contributions

1. Natural language explanations: Infinitely many sequences
    - Contrastive explanations: What explanations humans ask for and how they explain themselves
    - Finite set of contrastive templates : prompt PLMs to elicit contrastive explanations

2. Humans find them relevant, useful and easy to understand
    - Human judgement of grammaticality, relevance, factuality and usefulness

3. Models actually use them for prediction
    - Model incorporates contrastive explanations for commonsense reasoning
    - Manipulate contrastive explanations to quantify extent of usage by model.

Motivation: Humans Prefer Contrastive Explanations

Research in philosophy, psychology, and cognitive science (over 250 papers surveyed by Miller et al., 2019):
Explanations are contrastive: when people ask for an explanation of an event – **the fact** — they (sometimes implicitly) are asking for an explanation relative to some **contrast (foil)** case;

“Why P?” => “Why P rather than Q?”

Motivation: Contrastive Explanations are computationally efficient

Research in philosophy, psychology, and cognitive science (over 250 papers surveyed by Miller et al., 2019) shows that explanations are contrastive: when people ask for an explanation of an event – **the fact** — they (sometimes implicitly) are asking for an explanation relative (anchored) to some **contrast (foil)** case;

Contrastive explanation is answer to the question "Why P rather than Q?"

Contrastive Question: Why is it a crow and not a magpie?
Contrastive Explanation: Crows only have black feathers while magpies have white and black feathers

The crow's size, wing-span, eye-color etc are not important to this distinction.

Motivation: Humans Explain their decisions through contrast

Humans asked to explain ~100 examples containing (**fact** and **foil**)

Humans ***contrast*** answer choices (**fact** and **foil**) on distinguishing attributes that are ***relevant*** to the decision.

- 76% of Winograd Schema
- 64% of Physical Commonsense

i) I picked up a bag of peanuts and raisins for a snack. I wanted a sweeter snack out so I ate the __ for now.
Contrastive Expl. - Peanuts are salty while raisins tend to be sweet.

ii) The geese prefer to nest in the fields rather than the forests because in the __ predators are more hidden.
Contrastive Expl. - Forests are denser than fields

Key Observation : Recurring language patterns

Motivation: Why Contrastive Explanations

Social Attribution

Humans ask for (sometimes implicitly) contrastive explanations and are likely to use contrastive explanations when provided the fact and foil.

Computational benefits

Instead of exhaustively enlisting all reasons for the fact, contrastive explanations only explain why the fact is more likely than the foil.

Do PLMs contrast?

Do PLMs contrast **fact** and **foil** ? - Hard question

Models have billions of parameters that interact in complex ways

Knowledge about an entity, pair of entities is distributed

Do PLMs contrast?

Do PLMs contrast **fact** and **foil** ? - Hard question

Models have billions of parameters that interact in complex ways

Knowledge about an entity, pair of entities is distributed

Make PLMs **ELICIT** contrastive explanation explicitly

Provide the right interface (prompt) to PLM to extract **targeted** knowledge.

How do we know what language models know? Prompting PLMs

The knowledge contained in LMs is **probed** by providing a prompt, and letting the LM either

- generate the continuation of a prefix (e.g. “Barack Obama was born in _”)
- predict missing words in a cloze-style template (e.g., “Barack Obama is a _ by profession”)

	RoBERTa	Barack Obama was born in **Kenya** Barack Obama is a lawyer by profession
Barack Obama was born in _ Barack Obama is a _ by profession	GPT-2	Barack Obama was born in **Hawaii**
	T5-11B	Barack Obama was born in **Hawaii** Barack Obama is a lawyer by profession

Prompting to peek into PLMs

Model analysis/debugging : Can PLMs compare sizes or age, can they count

Probe name	Setup	Example	Human[1]
ALWAYS-NEVER	MC-MLM	*A chicken [MASK] has horns.* **A.** never **B.** rarely **C.** sometimes **D.** often **E.** always	91%
AGE COMPARISON	MC-MLM	*A 21 year old person is [MASK] than me in age, If I am a 35 year old person.* **A.** younger **B.** older	100%
OBJECTS COMPARISON	MC-MLM	*The size of a airplane is [MASK] than the size of a house .* **A.** larger **B.** smaller	100%
ANTONYM NEGATION	MC-MLM	*It was [MASK] hot, it was really cold .* **A.** not **B.** really	90%
PROPERTY CONJUNCTION	MC-QA	*What is usually located at hand and used for writing?* **A.** pen **B.** spoon **C.** computer	92%
TAXONOMY CONJUNCTION	MC-MLM	*A ferry and a floatplane are both a type of [MASK].* **A.** vehicle **B.** airplane **C.** boat	85%
ENCYC. COMPOSITION	MC-QA	*When did the band where Junior Cony played first form?* **A.** 1978 **B.** 1977 **C.** 1980	85%
MULTI-HOP COMPOSITION	MC-MLM	*When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest* **A.** second **B.** first **C.** third	100%

Table 1: Examples for our reasoning probes. We use two types of experimental setups, explained in §2. A. is the correct answer.

Talmor et al., 2019 [oLMpics -- On what Language Model Pre-training Captures]

Interpretability for PLMs that “solve” Commonsense Reasoning

Do PLMs contrast **fact** and **foil** ? - Hard question

Models have billions of parameters that interact in complex ways

Knowledge about an entity or word is distributed

Make PLMs **ELICIT** such contrastive knowledge explicitly.

Solution: Provide the **right interface (prompt)** to PLM to extract **targeted** knowledge.

Targeted Knowledge: Contrastive knowledge between **fact** and **foil**

Outline

Method

M1: Designing contrastive Prompts
M2: Prompting PLMs for contrastive explanations
M3: Using contrastive explanations in a downstream model for commonsense reasoning

Results

R1: Do contrastive explanations improve performance on commonsense tasks
R2: Do humans find contrastive explanations useful?
R3: Do models actually use explanations to solve the task?

Outline

Method

M1: Designing contrastive Prompts

M2: Prompting PLMs for contrastive explanations

M3: Using contrastive explanations in a downstream model for commonsense reasoning

Results

R1: Do contrastive explanations improve performance on commonsense tasks

R2: Do humans find contrastive explanations useful?

R3: Do models actually use explanations to solve the task?

M1: Designing contrastive prompts

3 In-house annotators asked to explain why one answer (FACT) is more likely than the other (FOIL) for 250 training instances.

Recurring patterns : P are more _ than Q, P have _ while Q have _

Dataset Instance	Human-Authored Contrastive Explanation
Winograd Schema 1. The party was more interesting and uplifing than the funeral because the __ was rigid. 2. The geese prefer to nest in the fields rather than the forests because in the _ predators are more hidden.	∘ Parties are for celebrating while funerals are for mourning ∘ People wear colorful clothes at parties and black at funerals ∘ Forests are dense while fields are sparse ∘ Forests have more predators than fields.

M1: Designing Contrastive Prompts

1. Manually examined ~250 explanations
2. Abstracted into templates containing at least two placeholders:
    - Fact
    - Foil
    - Property contrasted on
    - Eg. Peanuts are saltier than raisins: P is more _ than Q
3. Templates used > 10 times retained => ~50 templates
4. Coverage : Annotators used templates in over 82% cases for Winograd and PIQA

Prompt Pattern
Personal Characteristics
$\Longrightarrow$ P likes/likes to _ while Q likes/likes to _
P likes/likes to _ while Q does not like/like to _
P prefers/prefers to _ while Q prefers _
Q prefers _ while P does not prefer/prefer to _
Q thinks _ while P thinks/does not think _
Object Characteristics
P is taller/shorter/smaller/larger/slower/faster than Q
$\Longrightarrow$ P is/are _ while/but/however Q is/are _
Q has/have _ while/but/however P has/have _
P has/have more/less _ than Q
P is/are _ than Q
Spatial/Temporal Contrast
$\Longrightarrow$ P is inside/outside/above/below Q
_ is closer to P and farther away from Q
P is to the right/left of Q
Q takes longer to _ than P
Use cases and causes
P is used for _ Q
P is used to do Q _
$\Longrightarrow$ P is used for/to/in _ while Q is used for/to/in _
Q is used _ while P is used _
Q because _ while P because _
Q can cause _ while P results in _

Outline

Method

M1: Designing contrastive Prompts

M2: Prompting PLMs for contrastive explanations

M3: Using contrastive explanations in a downstream model for commonsense reasoning

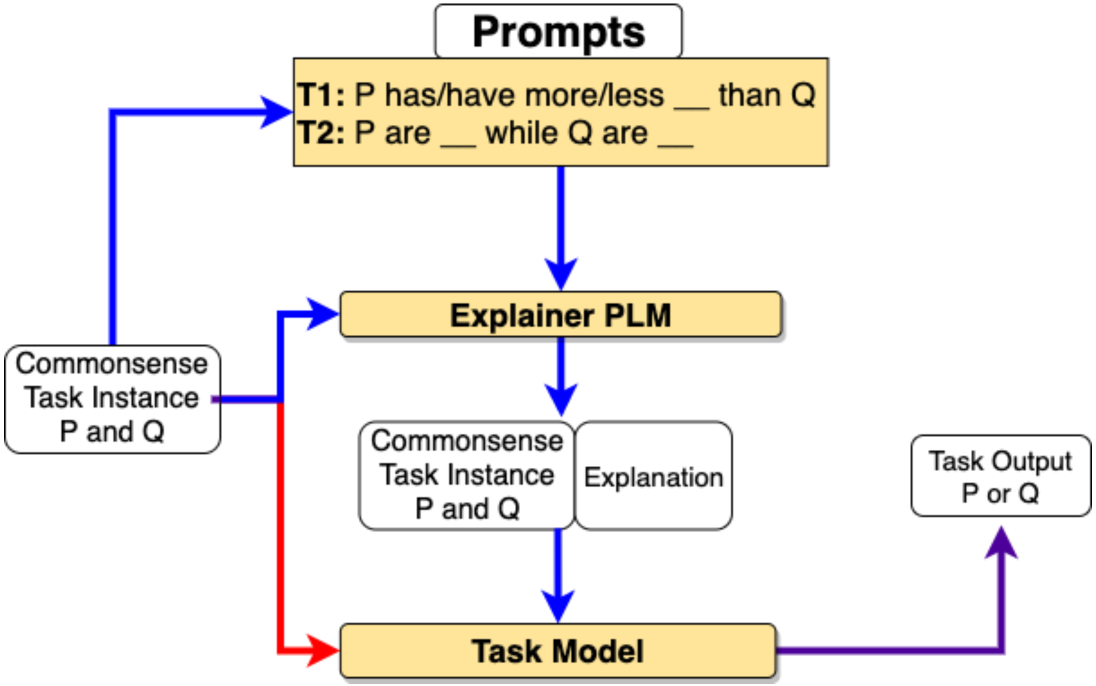
Results

R1: Do contrastive explanations improve performance on commonsense tasks

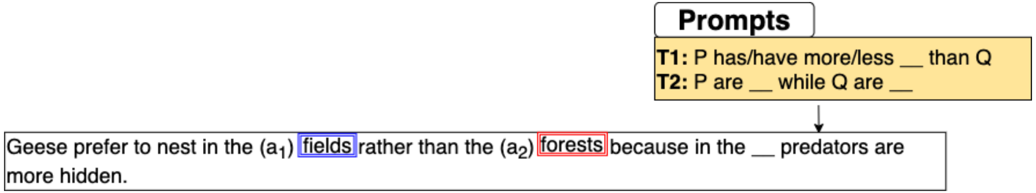
R2: Do humans find contrastive explanations useful?

R3: Do models actually use explanations to solve the task?

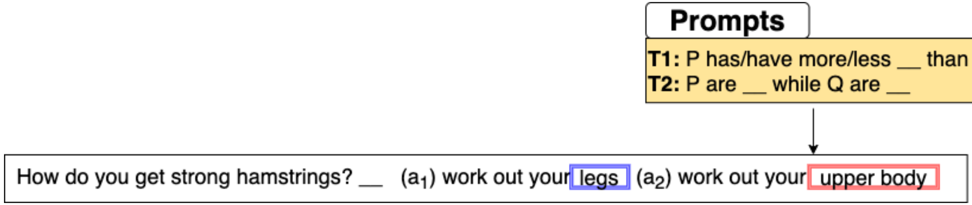
General Pipeline



M2: Prompting Contrastive Explanations



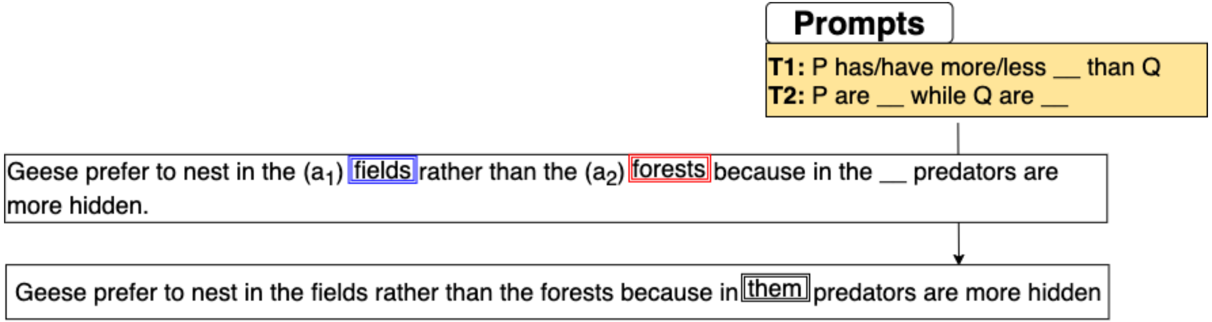
Winograd



PIQA

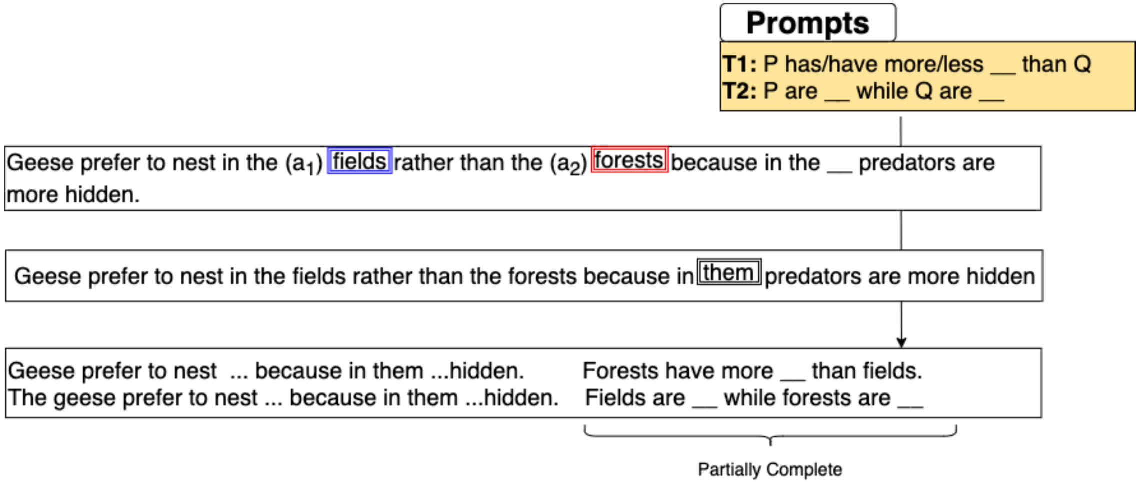
=> Identify **fact** and **foil** in the input context, which are typically two noun phrases surrounded by some context

M2: Prompting Contrastive Explanations



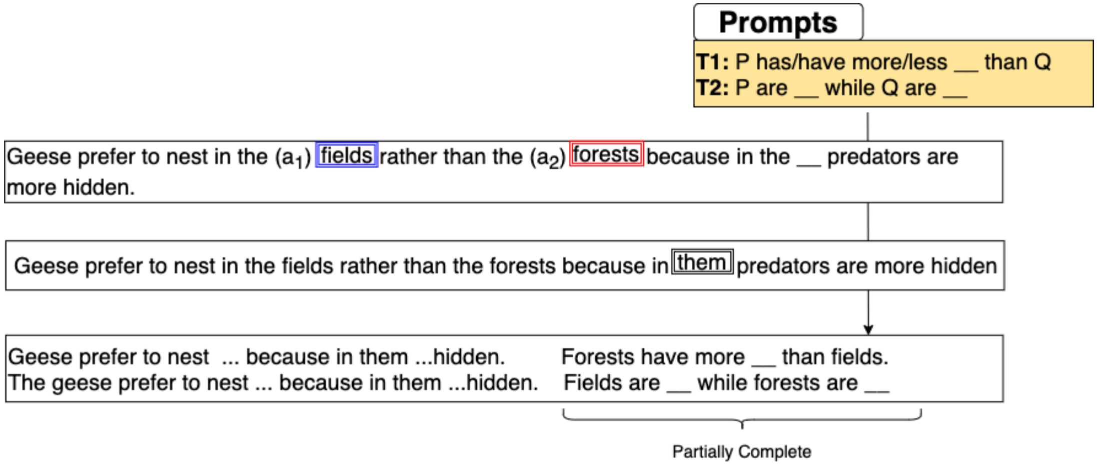
=> A neutral context *: A* complete sentence that contains fact and foil but no indication of the answer.

M2: Prompting Contrastive Explanations



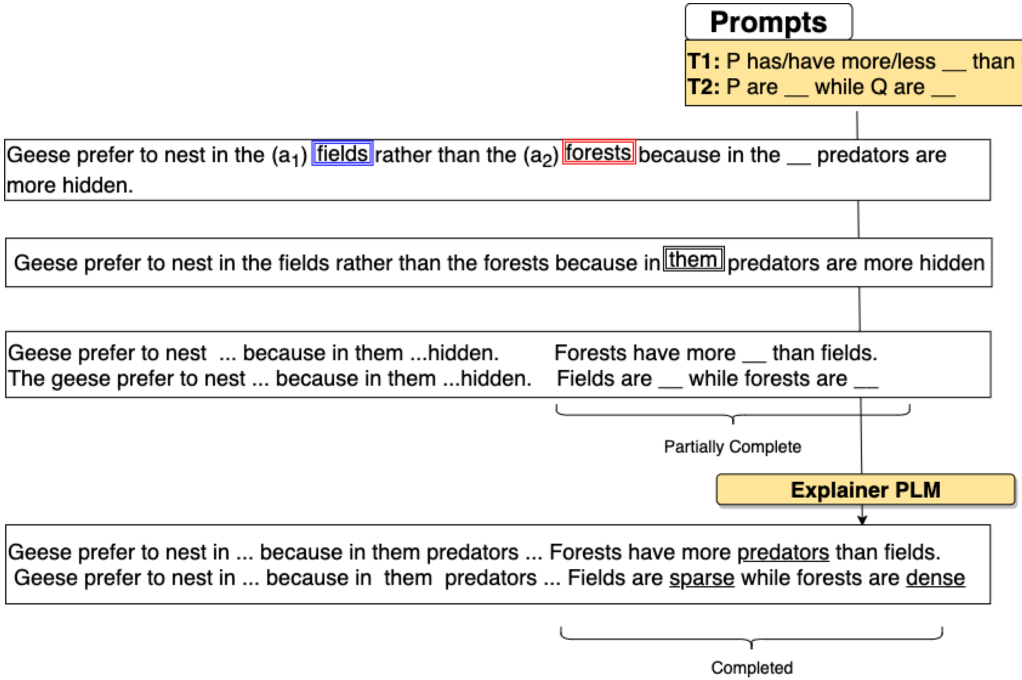
=> Initialize the template with **fact** and **foil.**

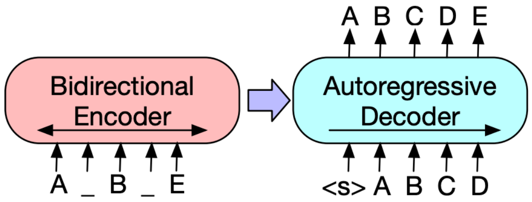
M2: Prompting Contrastive Explanations



=> The partially completed template (filled in with **fact** and **foil**) is appended to the neutral context

M2: Prompting Contrastive Explanations





The explainer PLM fills out the remaining portion of the template with contrastive knowledge that maybe embedded in its parameters. We get one explanation for every prompt.

Outline

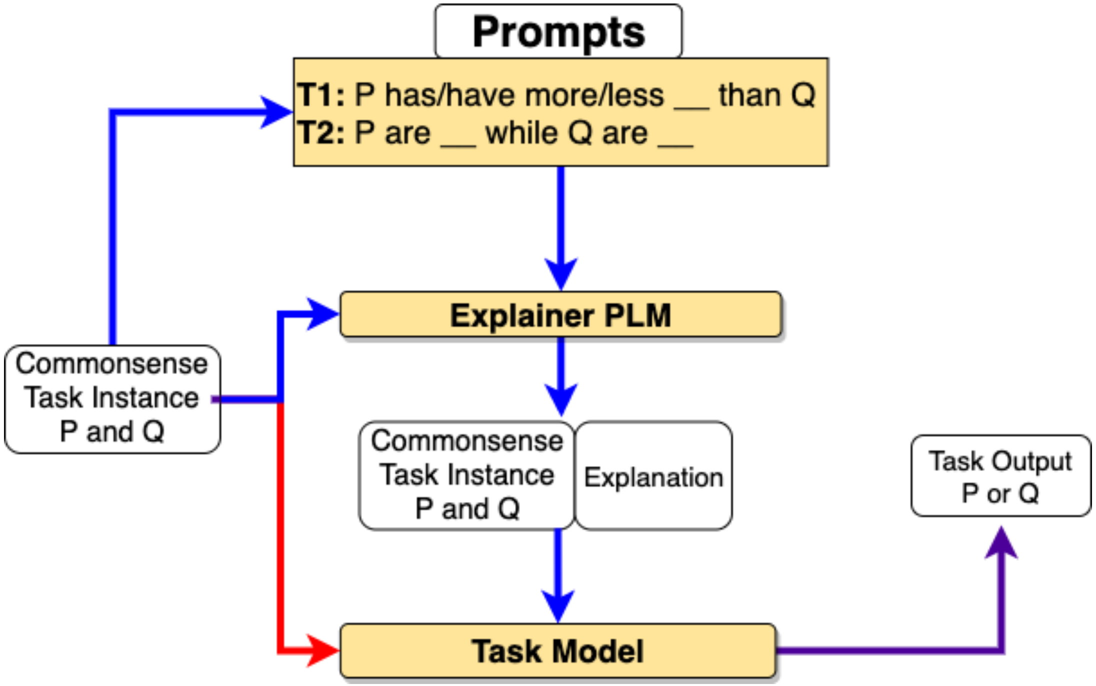
Method

M1: Designing contrastive Prompts
M2: Prompting PLMs for contrastive explanations
M3: **Using contrastive explanations in a downstream model for commonsense reasoning**

Results

R1: Do contrastive explanations improve performance on commonsense tasks
R2: Do humans find contrastive explanations useful?
R3: Do models actually use explanations to solve the task?

General Pipeline



M3: Zero-shot Model for Commonsense Reasoning

Transform into two *complete* sentences, that contain one of the answers
Language Model: Which alternative is more likely measured in terms of log-probability of the sentence.

```
The GPS and map helped me navigate home, I got lost when the _ got turned
upside down.
     (a) I got lost when the GPS got turned upside down.
     (b) I got lost when the map got turned upside down.
```

Transform into two possible sentences:

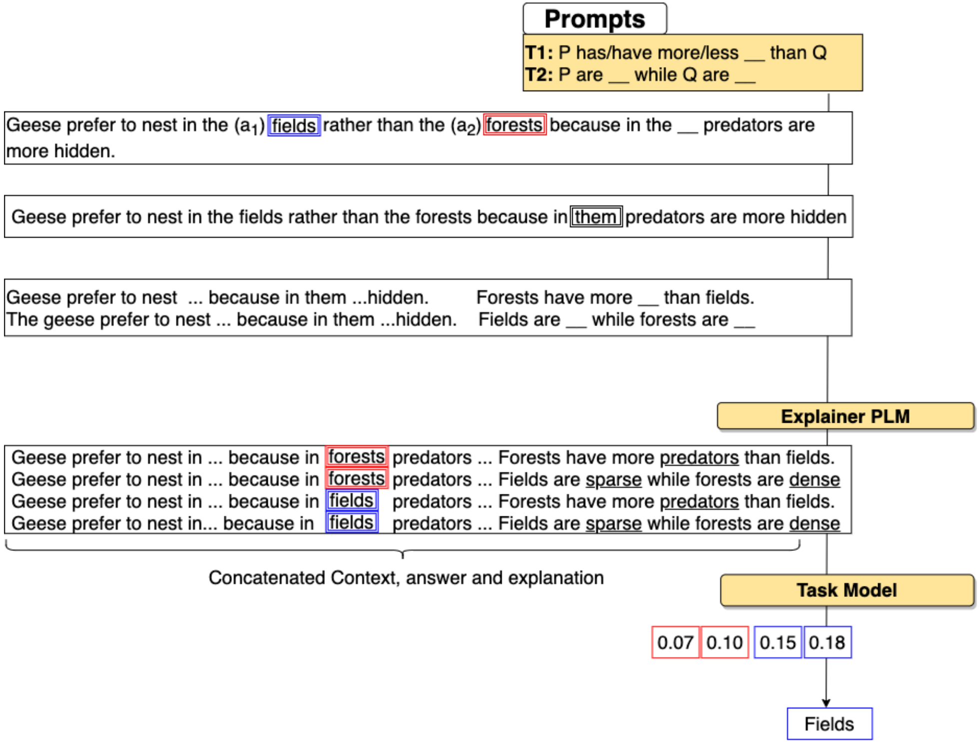
The GPS and map helped me navigate home, I got lost when the **GPS** got turned upside down. (0.056)

The GPS and map helped me navigate home, I got lost when the **map** got turned upside down. (0.078)

M3: Using Contrastive Knowledge

Generated explanation is concatenated with sentences containing one or the other answer.

The score for each answer is **aggregated** from different types of completed explanations.



Outline

Method

M1: Designing contrastive Prompts

M2: Prompting PLMs for contrastive explanations

M3: Using contrastive explanations in a downstream model for commonsense reasoning

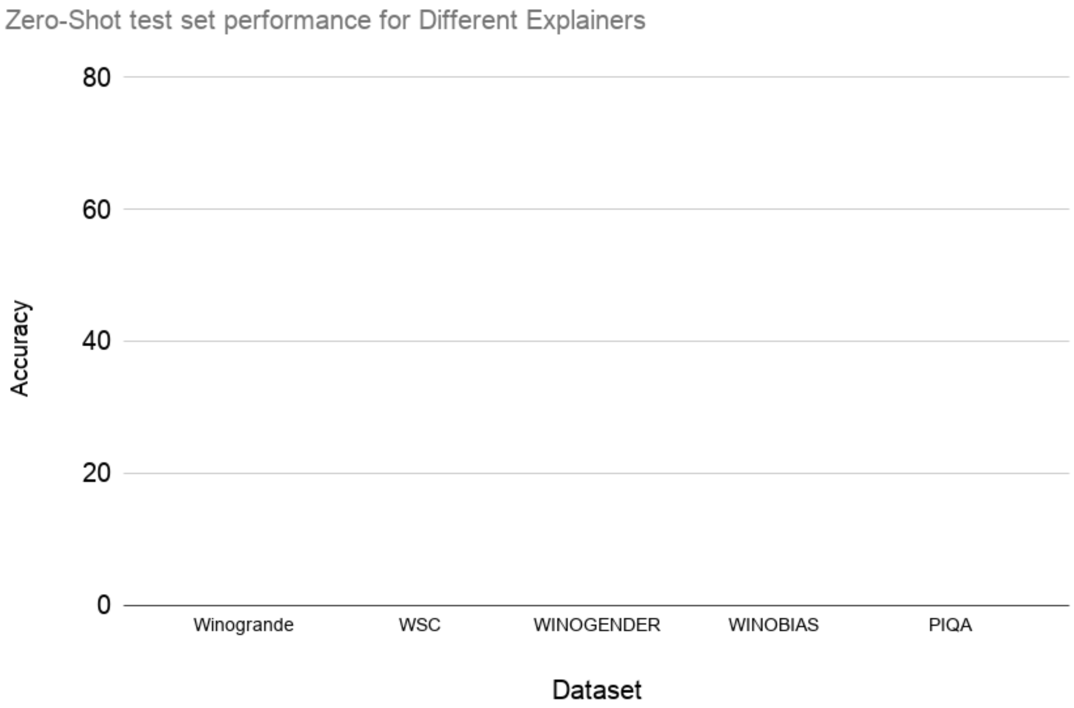
Results

R1: **Do contrastive explanations improve performance on commonsense tasks**

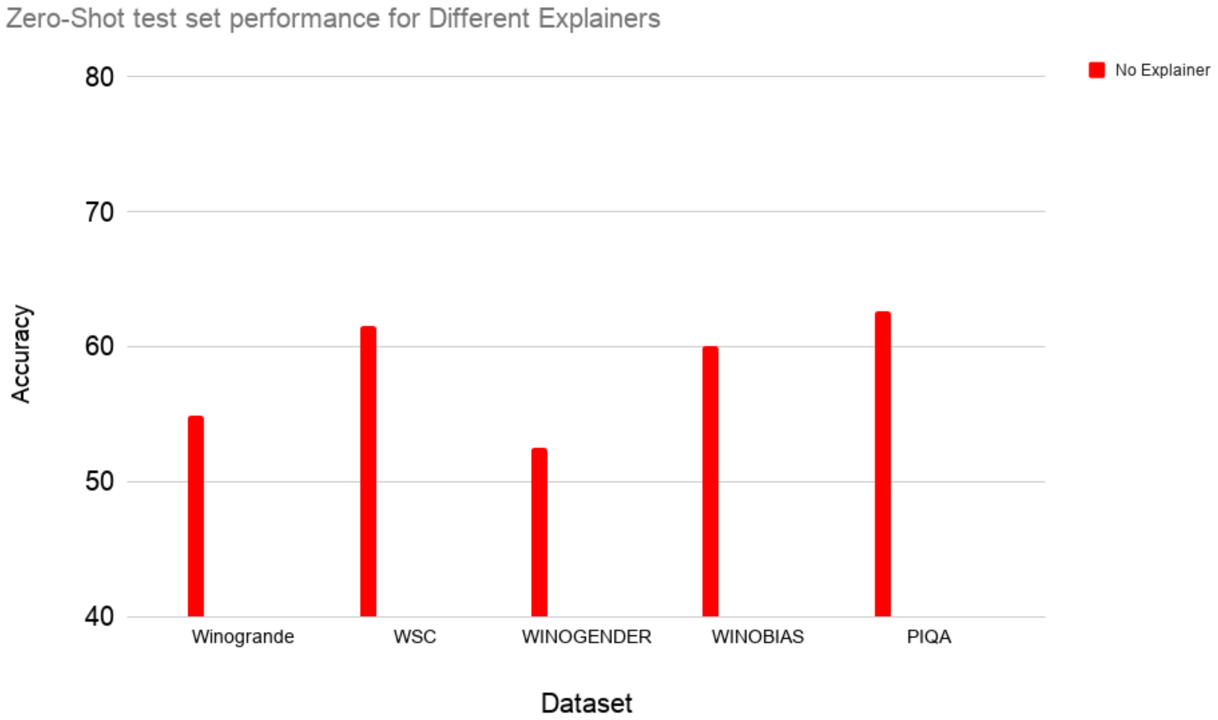
R2: Do humans find contrastive explanations useful?

R3: Do models actually use explanations to solve the task?

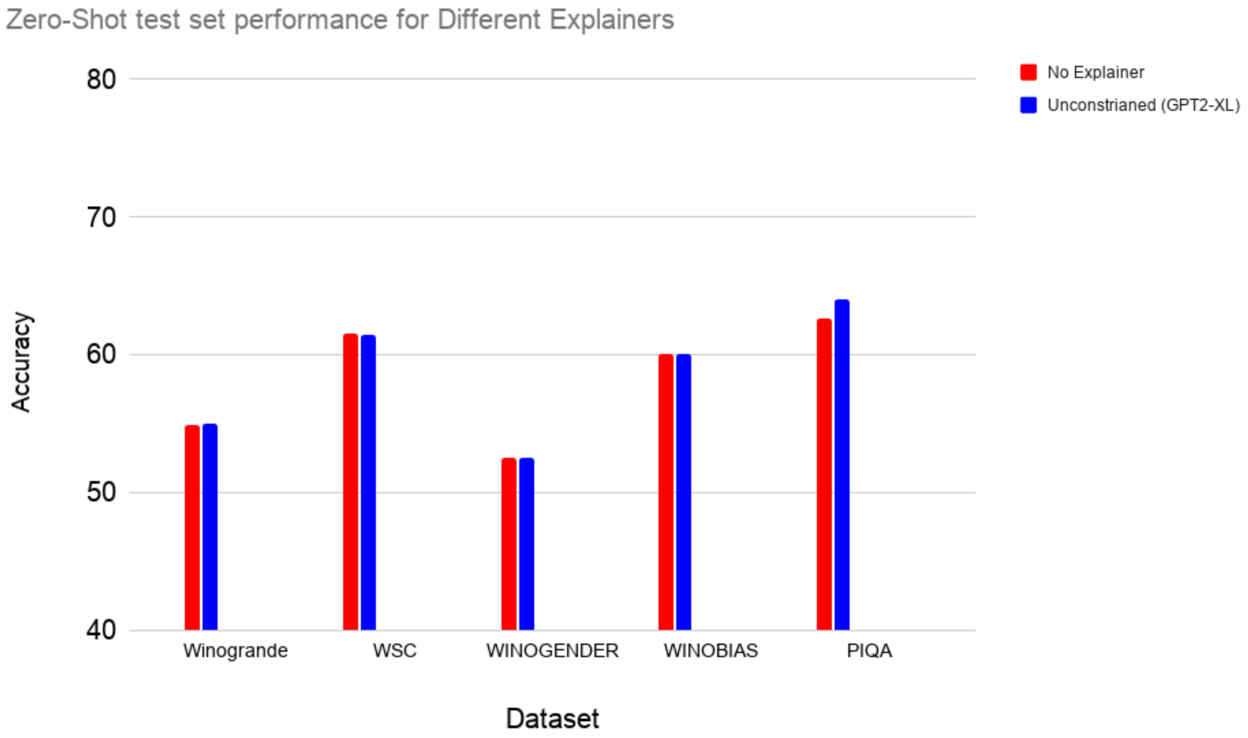
R1: Do contrastive explanations improve performance on commonsense tasks



R1: Do contrastive explanations improve performance on commonsense tasks



R1: Do contrastive explanations improve performance on commonsense tasks



Qualitative Examples

Example	Unconstrained	Self-Talk	Contrastive
(i) The GPS and map helped me navigate home. I got lost when the it got turned upside down.	because the GPS and map helped me navigate home.	What is going on here? The iphone app is not working.	The GPS can tell me where I am but the map can't. The GPS is right-side-up while the map is upside down
(ii) I helped my sister find her gold necklace. She couldn't wear her woven necklace to the ball because it was so casual.	She couldn't wear her woven necklace.	What are the properties of gold? The properties of gold are listed below.	Gold necklace is used for formal occasion while woven necklace is used for casual occasion.

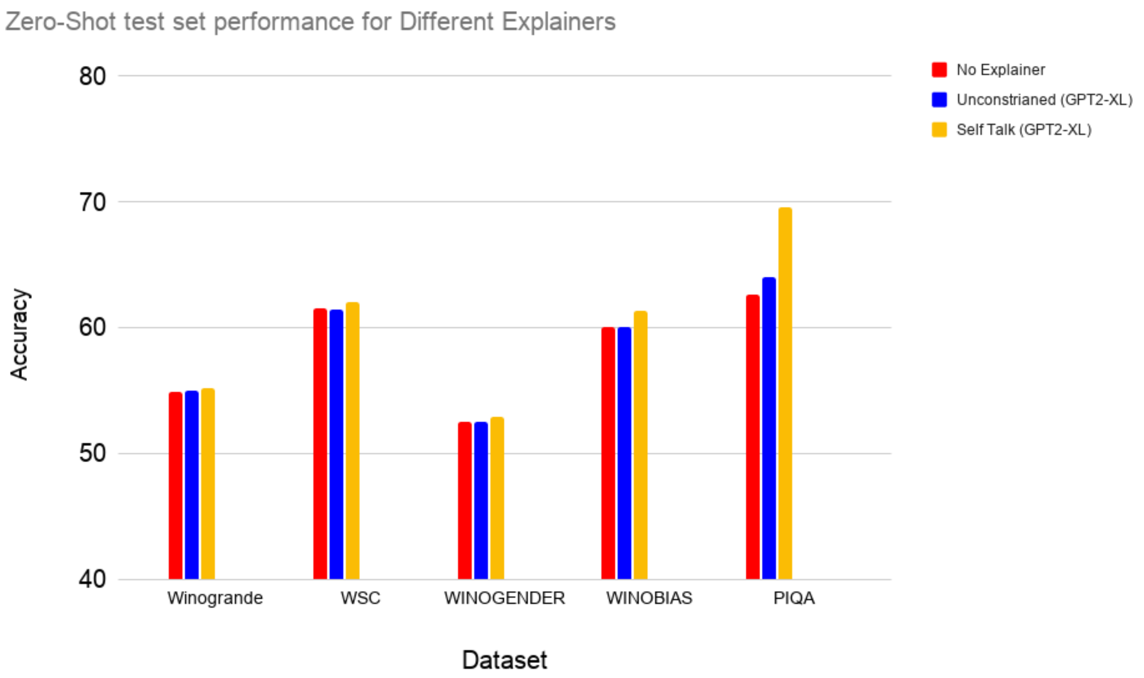
Table 6: Qualitative Examples on Winogrande where contrastive explanations (using T5-11B explainer) improve task performance over baselines.

Self-talk through clarification questions

Shwartz et. Al, 2020 [Unsupervised Commonsense Question Answering with Self-Talk]

Instance	Clarification
Irrelevant	
Q: how do you sit a baby in a restaurant? **Choices**: place them in a booster seat., place them on the table.	**Q:** What is the definition of "a good time"? **A:** The definition of "a good time" is not the same as what constitutes an acceptable meal.
Relevant	
The children were not vaccinated, which was fine with Betty but annoyed Mary. ___ believed they made kids autistic. **Choices**: Betty, Mary	**Q:** What does it mean to be "autistic"?" **A:** Be "autistic" means to have problems in social interaction and communication skills.

R1: Do contrastive explanations improve performance on commonsense tasks

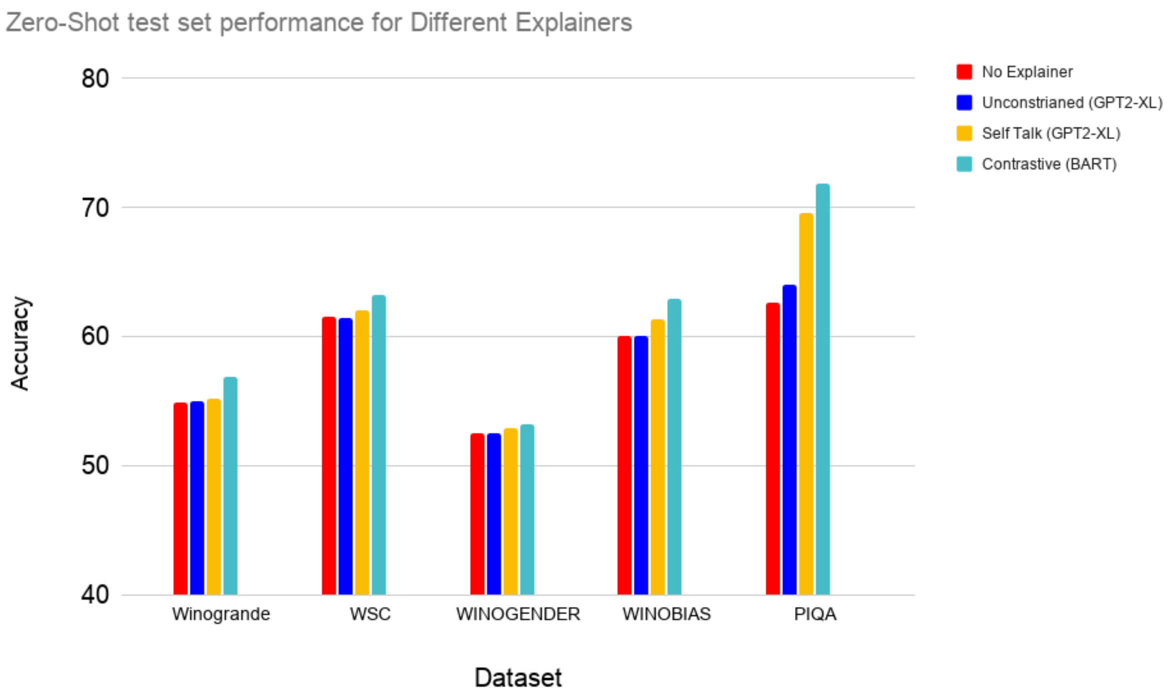


Qualitative Examples

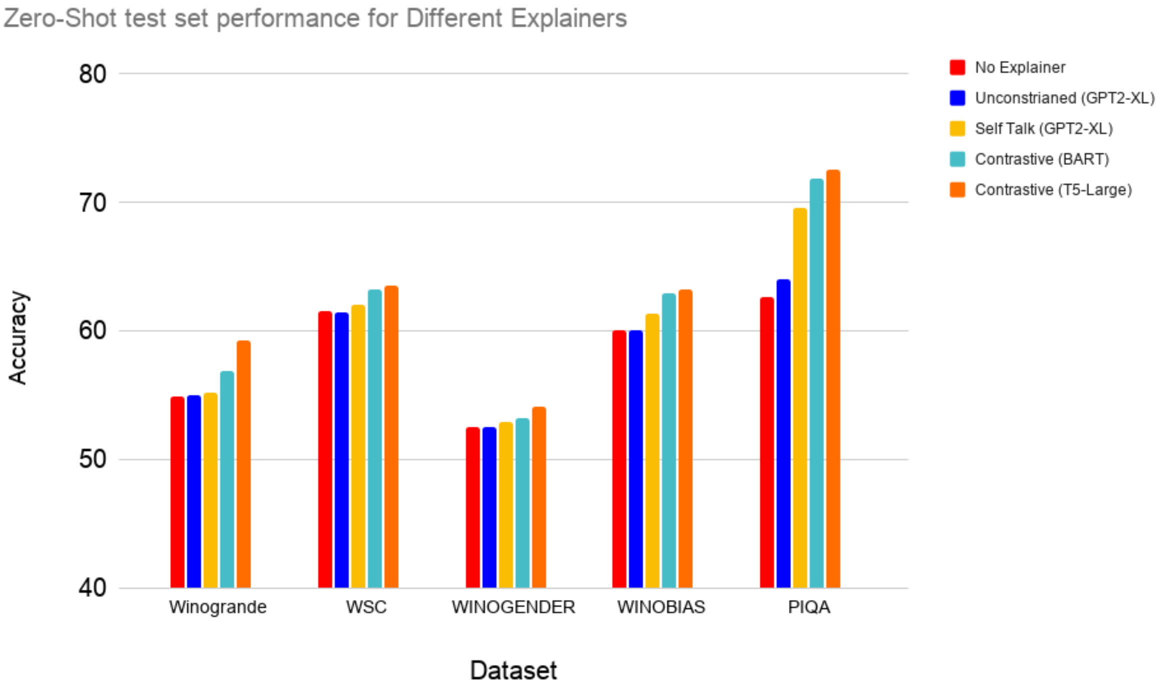
Example	Unconstrained	Self-Talk	Contrastive
(i) The GPS and map helped me navigate home. I got lost when the it got turned upside down.	because the GPS and map helped me navigate home.	What is going on here? The iphone app is not working.	The GPS can tell me where I am but the map can't. The GPS is right-side-up while the map is upside down
(ii) I helped my sister find her gold necklace. She couldn't wear her woven necklace to the ball because it was so casual.	She couldn't wear her woven necklace.	What are the properties of gold? The properties of gold are listed below.	Gold necklace is used for formal occasion while woven necklace is used for casual occasion.

Table 6: Qualitative Examples on Winogrande where contrastive explanations (using T5-11B explainer) improve task performance over baselines.

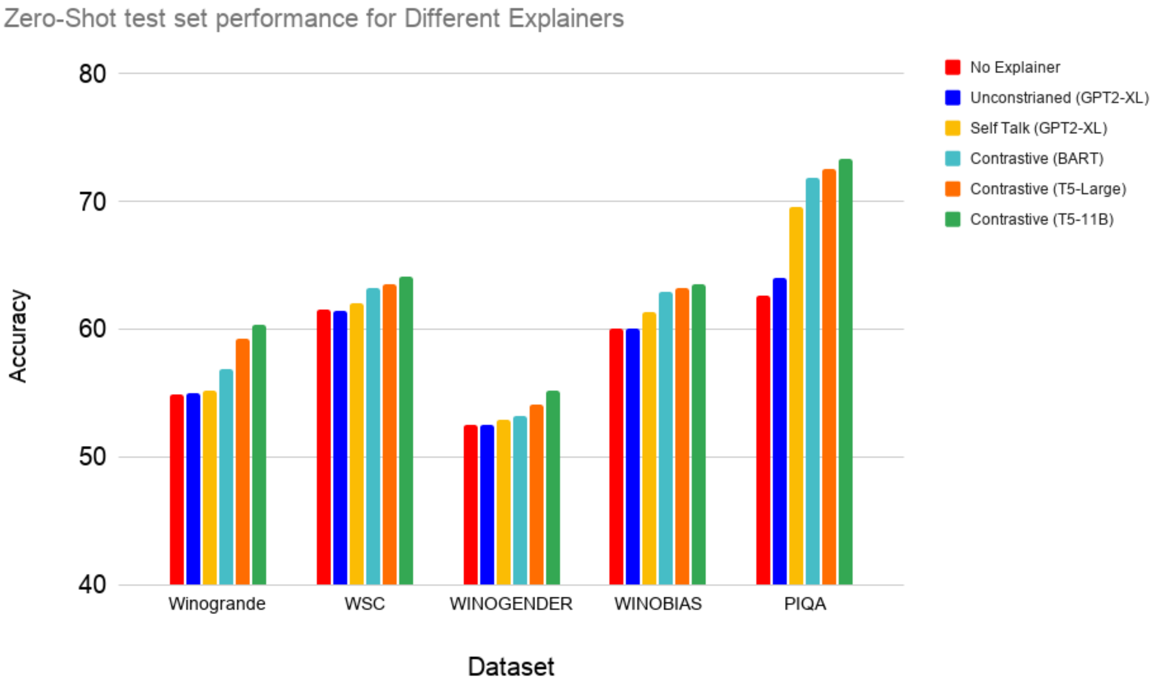
R1: Do contrastive explanations improve performance on commonsense tasks



R1: Do contrastive explanations improve performance on commonsense tasks



R1: Do contrastive explanations improve performance on commonsense tasks



Qualitative Examples

Example	Unconstrained	Self-Talk	Contrastive
(i) The GPS and map helped me navigate home. I got lost when the it got turned upside down.	because the GPS and map helped me navigate home.	What is going on here? The iphone app is not working.	The GPS can tell me where I am but the map can't. The GPS is right-side-up while the map is upside down
(ii) I helped my sister find her gold necklace. She couldn't wear her woven necklace to the ball because it was so casual.	She couldn't wear her woven necklace.	What are the properties of gold? The properties of gold are listed below.	Gold necklace is used for formal occasion while woven necklace is used for casual occasion.

Table 6: Qualitative Examples on Winogrande where contrastive explanations (using T5-11B explainer) improve task performance over baselines.

Outline

Method

M1: Designing contrastive Prompts
M2: Prompting PLMs for contrastive explanations
M3: Using contrastive explanations in a downstream model for commonsense reasoning

Results

R1: Do contrastive explanations improve performance on commonsense tasks?
R2: **Do humans find contrastive explanations useful?**
R3: Do models actually use explanations to solve the task?

R2 : Do humans find contrastive explanations useful?

AMT workers are asked to qualitatively judge 50 explanations

- Along 4 dimensions
- Independently for Self-talk and contrastive examples

Metric	Self-Talk(Reported)		Self-Talk		Contrastive	
	WGRD	PIQA	WGRD	PIQA	WGRD	PIQA
Relevant	68	60	70.4	61.7	73.1	70.7
Factual	46	42	40.8	38.8	43.0	39.4
Helpful	24	26	22.5	27.7	42.8	32.8
Grammatical	87.2	87.2	87.5	87.5	83.5	83.5

Outline

Method

M1: Designing contrastive Prompts
M2: Prompting PLMs for contrastive explanations
M3: Using contrastive explanations in a downstream model for commonsense reasoning

Results

R1: Do contrastive explanations improve performance on commonsense tasks?
R2: Do humans find contrastive explanations useful?
R3: **Do models actually use explanations to solve the task?**

R3 : Do models actually use explanations?

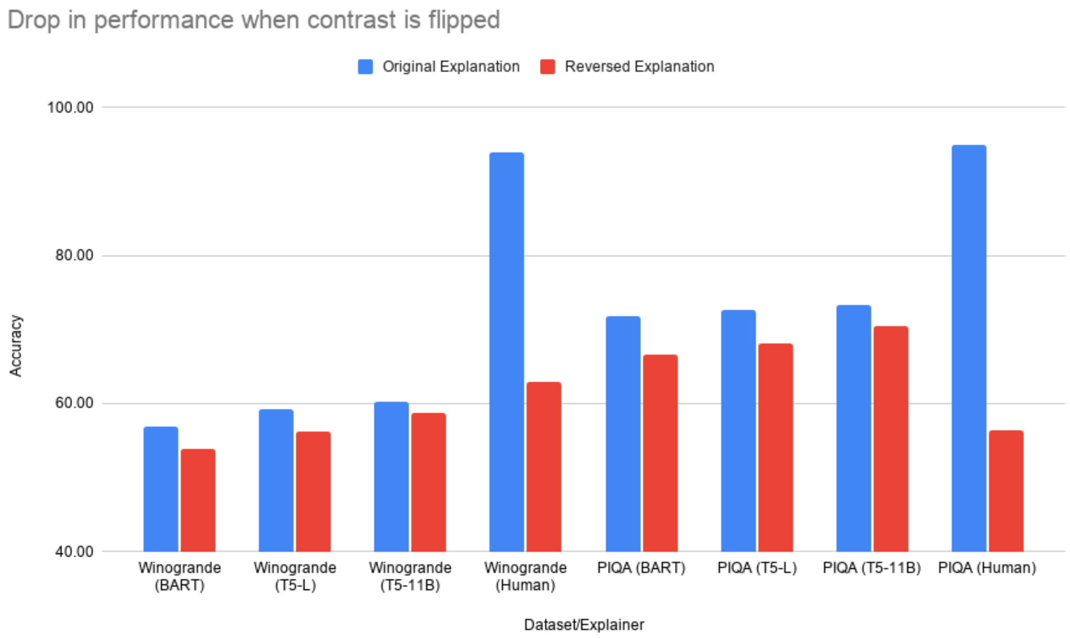
```
The GPS and map helped me navigate home, I got lost when the _ got turned upside
down.
     (a) I got lost when the GPS got turned upside down.
     (b) I got lost when the map got turned upside down.
GPS is fixed to the dashboard while a map can be moved freely


Reversed Explanation: Map is fixed to the dashboard while the GPS can be
moved freely
```

Expected Behavior : Task Model decision should ***flip*** if it relies on the explanation

The model should quantify the degree to which the explanation it provides is actually us
for prediction - degree of flip

R3 : Do models actually use explanations?



What if only the explainer sees FACT and FOIL

- X : Geese prefer to rest in fields rather than forests, because in _ predators are more hidden.
- Y : forests
- X’: Geese prefer to rest in A rather than B, because in _ predators are more hidden.
- e : Forests are denser than Fields
- e’ : B are denser than A

$$P(Y \mid X) := P_{task}(Y \mid X').$$

What if only the explainer sees FACT and FOIL

- X : Geese prefer to rest in fields rather than forests, because in _ predators are more hidden.
- Y : forests
- X’: Geese prefer to rest in A rather than B, because in _ predators are more hidden.
- e : Forests are denser than Fields
- e’ : B are denser than A

$$P(Y \mid X) := \sum_{e} P_{task}(Y \mid X, e) P_{expln}(e \mid X).$$

What if only the explainer sees FACT and FOIL

- X : Geese prefer to rest in fields rather than forests, because in _ predators are more hidden.
- Y : forests
- X': Geese prefer to rest in A rather than B, because in _ predators are more hidden.
- e : Forests are denser than Fields
- e' : B are denser than A

Input to task model would be : Geese prefer to rest in A rather than B, because in _ predators are more hidden. B are denser than A

- Non-abstracted explanation model $P_{expln}(E \mid X)$
- Abstracted decision model $P_{task}(Y \mid X', E)$

$$P(Y \mid X) := \sum_{e} P_{task}(Y \mid X', e') P_{expln}(e \mid X).$$

What if only the explainer sees FACT and FOIL

Input	WGRD	
Fully abstracted	63.2	$P(Y \mid X) := P_{task}(Y \mid X')$.
Abst. after expl.	70.4	$P(Y \mid X) := \sum P_{task}(Y \mid X', e')P_{expln}(e \mid X)$.
No abstraction	79.1	$P(Y \mid X) := \sum_{e} P_{task}(Y \mid X, e)P_{expln}(e \mid X)$.

Conclusion

- Contrastive explanations have social and computational significance
- Custom prompts are designed to ELICIT contrastive knowledge from large pre-trained models.
- Elicited explanation is found to be useful for the model and more meaningful to humans.
- The unique form of contrastive explanations allows us to manipulate the explanation to debug model behavior.

Future Work

- Implicit foils and multiple-choice questions with more than one foils
- How and where information is actually stored in parameters
- Techniques to isolate the importance of the faithfulness of the model to generated explanation.

Thank you

Limitations

- Implicit Foils
  - ❖ Choice of foil selection is challenging
  - ❖ Faithfulness - information encoded in choice
- Knowledge in Task Model
  - ❖ Can learn to ignore the explanation

Generalizability of Templates

- Commonsense QA (Talmor et al. 2019)

Model	Dev Accuracy	Test Accuracy
Random	20.0	20.0
Baseline	36.4	37.2
Self talk	32.4	26.9
Baral et. al. (ext. sources that relies on conceptnet)	38.2	38.8
Ours (Vote)	37.1	38.4
Ours (Max Margin)	36.5	38.1

Use-Cases

- Ambiguous answers to questions