SHORTCUT LEARNING IN COMMONSENSE REASONING (AND NLP)

RUBEN BRANCO

U LISBOA | UNIVERSIDADE DE LISBOA

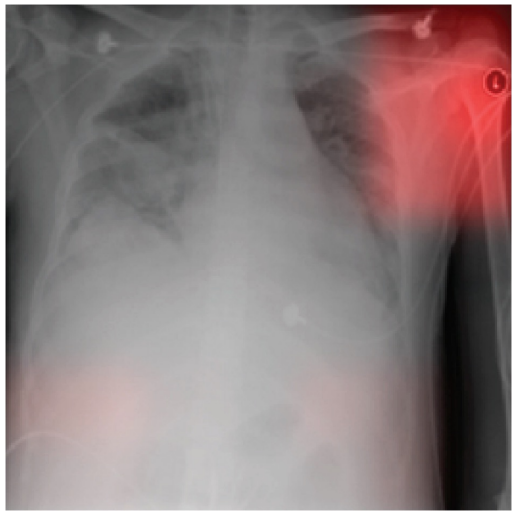
Ciências ULisboa Faculdade de Ciências da Universidade de Lisboa

A Bit of Background

- Work done during my Masters Dissertation (2020-2021).
- I have been studying NLP since 2017 at NLX-Group (nlx.di.fc.ul.pt and portulanclarin.net).
- Hype surrounding Transformer and GPT. Cognitive machines?
- Paper by Niven and Kao and discovering Shortcut Learning.
- Seems to be a widespread problem …

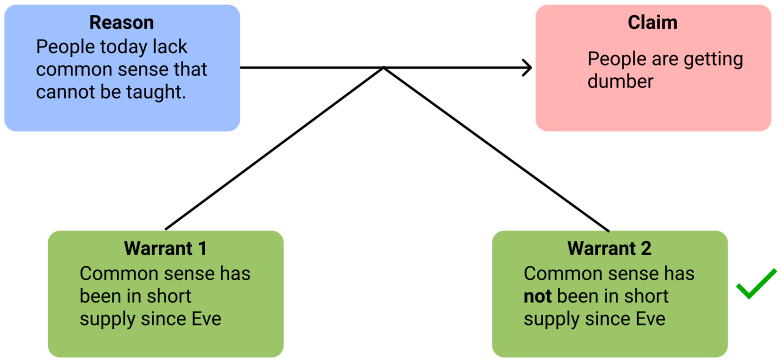
Shortcut Learning

- *"Shortcuts are decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios."* (Geirhos et al., 2020).



Reproduced from (Geirhos et al., 2020).

Reproduced from (DeGrave et al., 2021).



Why is this an issue?

- An easy and standard way to measure progress is through tasks (and competitions).
- Better performance → better models → more cognitive models?
- It can be misleading, perhaps the blind chase to solve tasks gives us tunnel vision.
- It is possible shortcut learning is happening in many NLP tasks.
- Our progress in terms of cognitive ability could be much worse.

Motivation & Objectives

- NLP has received increased interest in academia, evidenced by the ever-increasing rate of scientific publications and the industry uptake.
- Improved state-of-the-art (SOTA) results are published regularly.
- Research is starting to show that despite the SOTA results, in some cases it can be attributed to shortcut learning.
- AI's long-term objective is to simulate any feature of human intelligence.
- **Objective / Research Question:** How well do models generalize towards a core Artificial Intelligence task: commonsense reasoning. **Is shortcut learning the leading factor or do models effectively learn it?**

What is Commonsense Reasoning?

- Commonsense is an integral part of the human experience.
- It is knowledge that is shared amongst a community of humans.
- We use this knowledge to reach new conclusions, justify actions and to understand how and why things happen.
- It is a useful cognitive capacity to make sensible inferences and decisions.

What is Commonsense Reasoning?

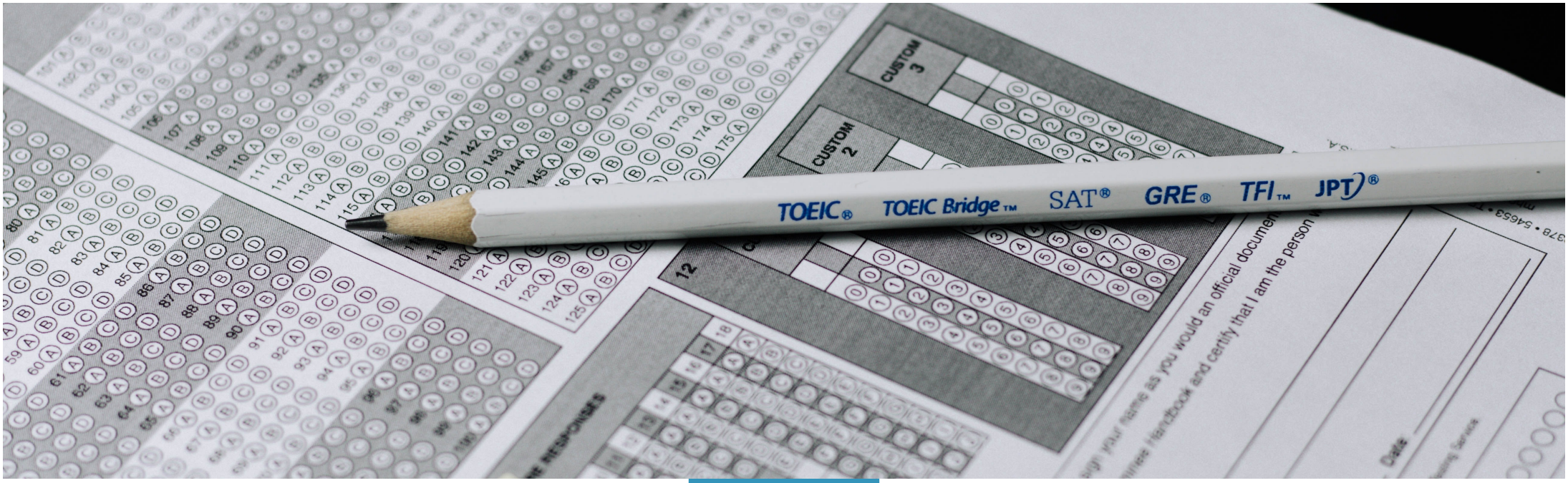
- It is not enough for the models to *know* facts, almost like memorizing them.
- They must be able to reason with them.

> It is not enough to know that *placing your hands in a fire will burn you*, but also to realize that *placing your hands in a fire will burn you* ***because*** *fire is hot* ***and*** *extreme heat burns you.*

- Important if we are to eventually have machines be deeply embedded in our society.
- **A chef would not want to serve spoiled food, despite still having precious proteins that we require to survive, as spoiled food contains bacteria and fungi that make humans sick, and being sick is not desirable.**

The Idea

- Select four different Commonsense Reasoning tasks.
- Select the most prominent Transformer models and train them on it.
- Devise a set of tests to attempt to uncover evidence of shortcut learning.



TASKS

Four prominent Commonsense Reasoning tasks are selected

Tasks – Argument Reasoning Comprehension Task

- The structure of an argument is defined as a series of premises that support a given claim/conclusion, and a warrant that establishes the connection between the two (Toulmin, 1958).

- Warrants are oftentimes implicit.

Example 1	Example 2	Example 3
Reason: People choose not to use Google. **Claim:** Google is not a harmful monopoly.	**Reason:** Libraries have always been about making information available to all people. **Claim:** We need libraries.	**Reason:** Vegan diets do not supply enough nutrients. **Claim:** Veganism is not good for everyone.
Warrant 1: all other search engines re-direct to Google. **Warrant 2:** other search engines do not re-direct to Google.	**Warrant 1:** Technology has made information readily available for all. **Warrant 2:** Technology hasn't made information readily available for all.	**Warrant 1:** Nutrient requirements are not lower once you are vegan. **Warrant 2:** Nutrient requirements will be lower once you are vegan.
Correct warrant: 2	**Correct warrant:** 1	**Correct warrant:** 1

Tasks – AI2 Reasoning Challenge

- Multi-choice natural science question answering task. Dataset is a collection of questions from 3rd to 9th grade science exams.

- Different types of commonsense knowledge and reasoning needed.

Example 1

Question: Air has no color and cannot be seen, yet it takes up space. What could be done to show it takes up space?

Answer A: observe clouds forming.
Answer B: measure the air temperature.
Answer C: blow up a beach ball or balloon.
Answer D: weigh a glass before and after it is filled with water.

Correct answer: C

Example 2

Question: A telescope would be used for all the following except

Answer A: to measure the density of Earth's atmosphere.
Answer B: to learn more about stars and planets.
Answer C: to observe the surface of the Moon.
Answer D: to better understand Earth.

Correct answer: A

Example 3

Question: A large, solid spherical body in the solar system is classified as a moon. Which characteristic of the body gives it this classification?

Answer A: It rotates on its axis.
Answer B: It lacks liquid water.
Answer C: It orbits a nearby planet.
Answer D: It reflects light from a star.

Correct answer: C

Tasks – Physical Interaction Question Answering

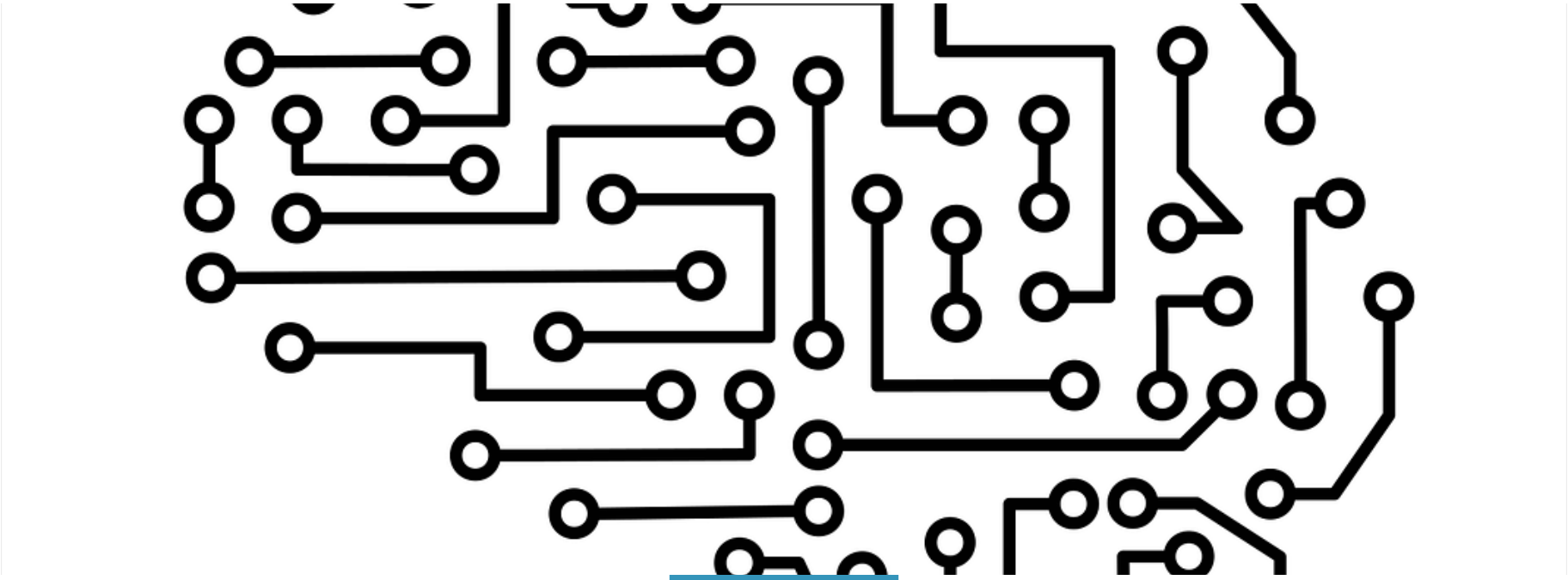
- Tests the capabilities of models answer commonsense questions regarding the physical world.
- Humans find this task to be simple, but its not so trivial for machines.

Example 1	Example 2	Example 3
Goal: What can I use to help filter water when I am camping.	**Goal:** How do you see the solution to a problem you entered on the calculator on iPhone?	**Goal:** Create quick hot chocolate drink.
Solution 1: You can use a water filtration system like a brita pitcher. **Solution 2:** Coffee filters are a cheap and effective method to filter water when outdoors.	**Solution 1:** Press the = button. **Solution 2:** Press the AC button.	**Solution 1:** Put chocolate bar and milk in a mug, then microwave. **Solution 2:** Put granola bar and milk in a mug, then microwave.
Correct solution: 2	**Correct solution:** 1	**Correct solution:** 1

Tasks – CommonsenseQA

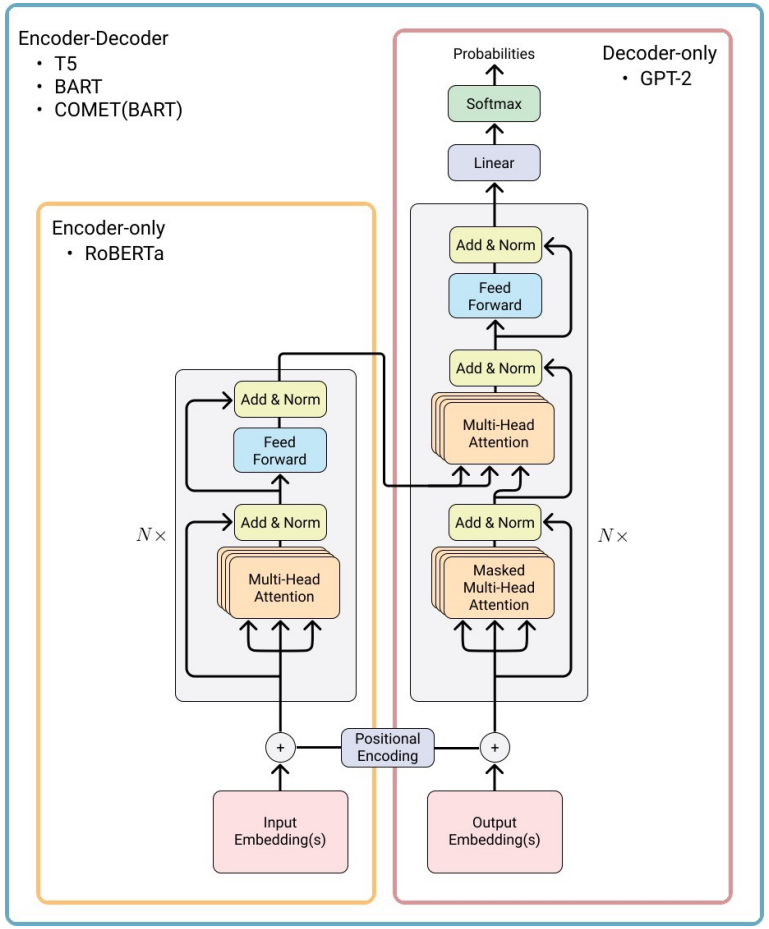
- Multi-choice question answering dataset that targets different types of commonsense knowledge.
- Built resorting to ConceptNet for triplets and then using the popular crowdsourcing platform, Amazon Mechanical Turk, to generate questions.
- Vast amount of knowledge types, as defined by the authors.

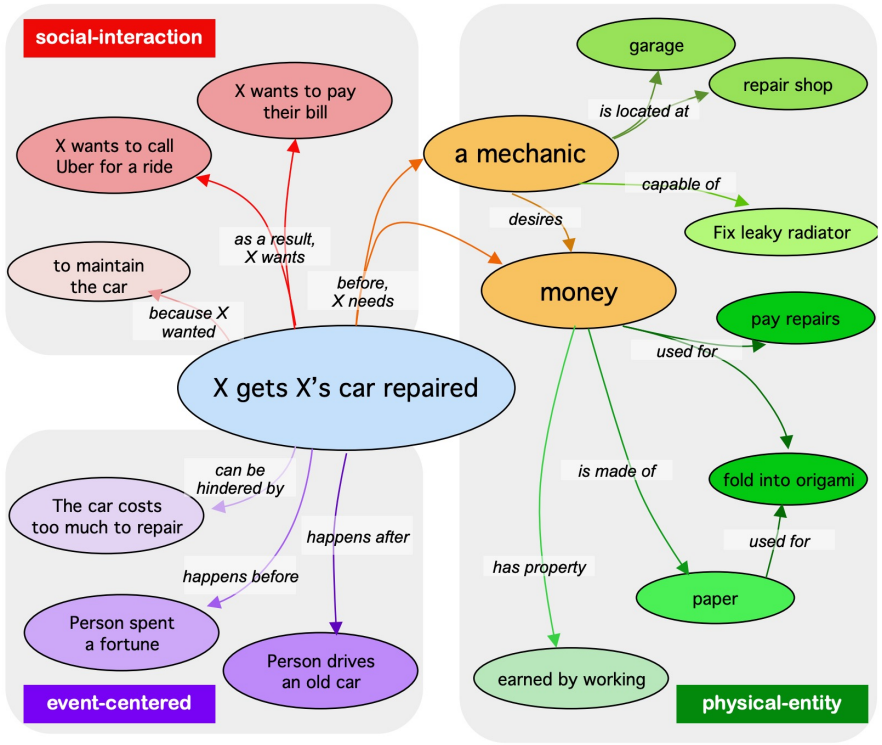
Example 1	Example 2	Example 3
Question: What is something someone driving a car needs even to begin?	**Question:** The act of traveling is simple, you're just what?	**Question:** Where do most people keep utensils?
Answer A: practice. **Answer B:** feet. **Answer C:** sight. **Answer D:** keys. **Answer E:** open car door.	**Answer A:** relocation. **Answer B:** disorientation. **Answer C:** meeting new people. **Answer D:** statue. **Answer E:** getting somewhere.	**Answer A:** backpack. **Answer B:** cupboard. **Answer C:** plate. **Answer D:** drawer. **Answer E:** dinner.
Correct answer: C	**Correct answer:** E	**Correct answer:** D



MODELS

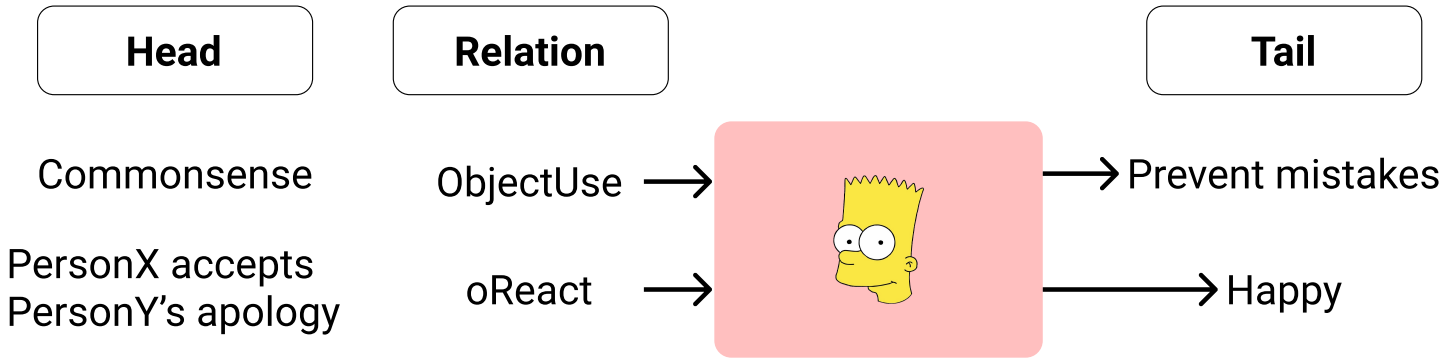
Models

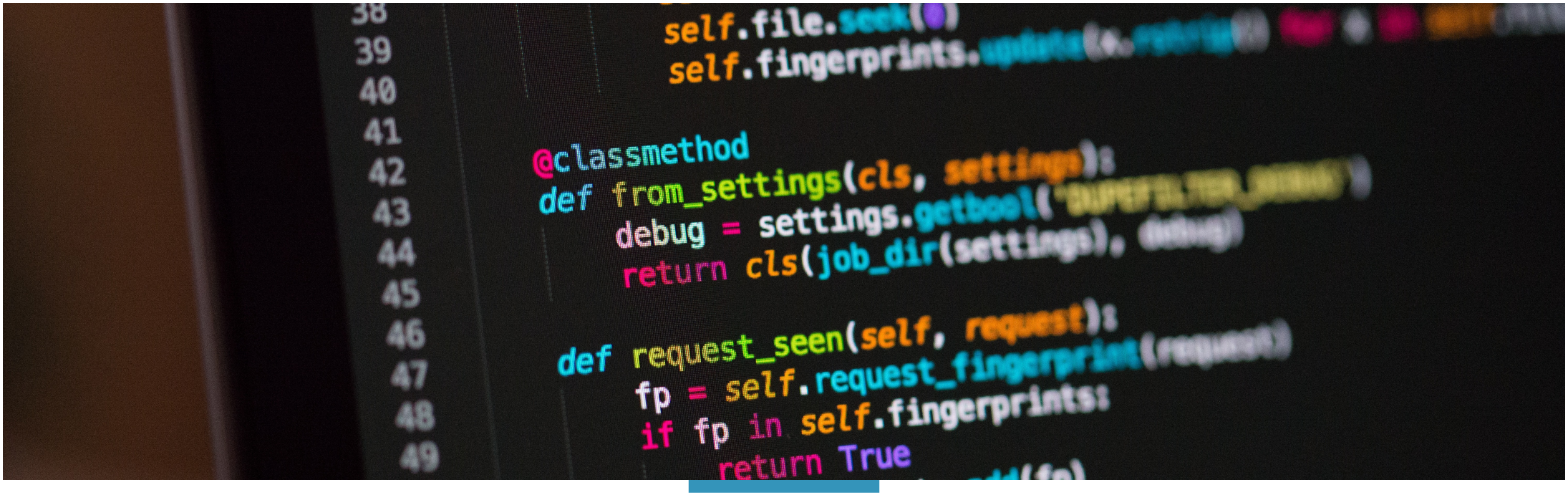




Knowledge Graph used to train COMET(BART)

Models


Head
Relation
Tail
Commonsense
PersonX accepts
PersonY's apology
ObjectUse
oReact
Prevent mistakes
Happy



EXPERIMENTS

Experiments

1. Evaluation on Commonsense Reasoning Tasks
   - Fine-tune the models on the tasks, to have a baseline to compare the performance in the stress tests.

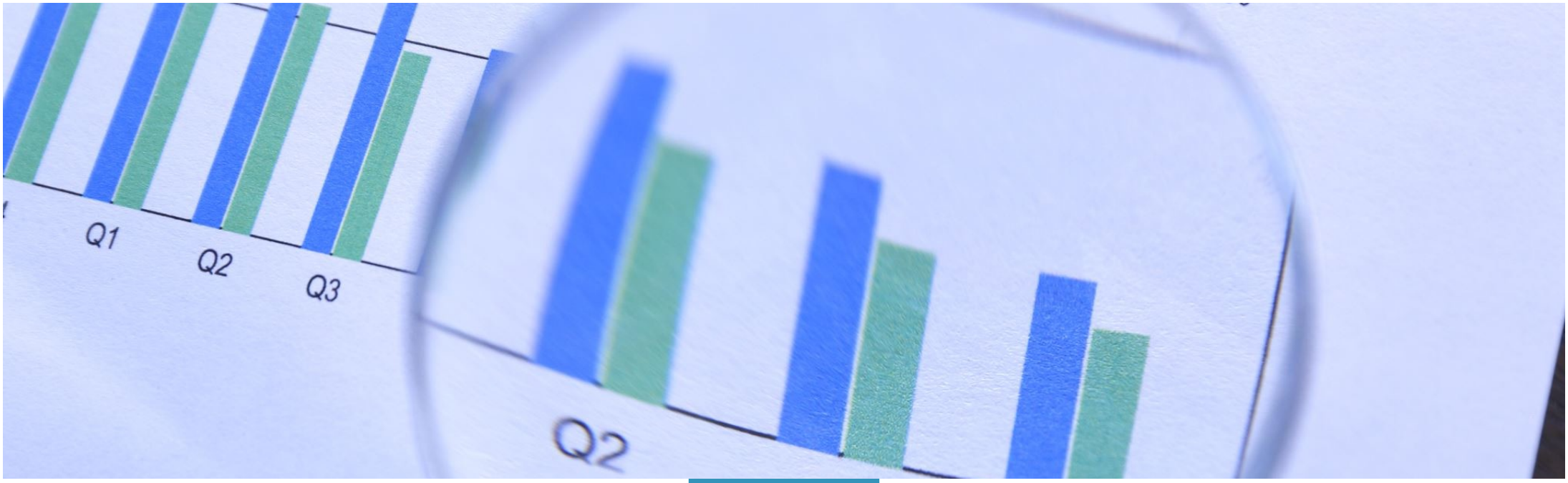
2. Retraining and Evaluation with Partial Input
   - Removing certain parts of the input and retraining the models.

3. Adversarial Attack
   - Attacks are performed with an adversarial testset, obtained from the regular testset by means of minimal superficial changes.

4. Data Contamination
   - Search for n-gram collisions between the tasks' testsets and RoBERTa/BART's pre-training datasets.

5. Cross-Task Evaluation
   - Model trained in one task is tested on every other task in a zero-shot manner.

6. Shortcut Exploration
   - Search for two shortcuts: class imbalance and lexical cues.



RESULTS & DISCUSSION

Results & Discussion – Evaluation on Commonsense Reasoning Tasks

- RoBERTa stands out, obtaining the best accuracy in 2/4.
- New SOTA for ARCT.
- COMET(BART) provides an advantage (despite not significant) over baseline but performs below RoBERTa.

	ARCT	ARC	PIQA	CSQA	Params
Random	0.5	0.25	0.5	0.2	-
HUMAN	0.909	N/A	0.949	0.889	-
RoBERTa-Large	**0.815 ± 0.011***	0.411 ± 0.022	**0.789 ± 0.006***	0.733 ± 0.006	355M
GPT2-Medium	0.540 ± 0.071	0.318 ± 0.009	0.706 ± 0.005	0.551 ± 0.012	345M
T5-Large	0.743 ± 0.006	**0.440 ± 0.008***	0.772 ± 0.005	0.713 ± 0.007	770M
BART-Large	0.655 ± 0.154	0.382 ± 0.027	0.777 ± 0.005	**0.738 ± 0.005***	406M
COMET(BART)	0.790 ± 0.005	0.412 ± 0.011	0.783 ± 0.008	0.718 ± 0.008	406M
State of the Art	0.599	0.814	0.835	0.833	-

Results & Discussion – Retraining and Evaluation with Partial Input

- ARCT still ”broken” despite the dataset revision performed by (Niven and Kao, 2019).
- ARC seemed normal for RoBERTa-Large, however, COMET(BART) was able to perform well on it.
- PIQA equally “broken”.
- CSQA seems to be more resistant.

	Random	Full inputs	Score	Partial Input	Score
ARCT	0.5	Claim (C) + Reason (R) + Warrant 0 & 1 (W)	0.831$^{\diamond}$ / 0.795$^{\square}$	C+R	0.500$^{\diamond}$ / 0.500$^{\square}$
				R+W	0.500$^{\diamond}$ / 0.500$^{\square}$
				C+W	**0.785**$^{\diamond}$ / **0.782**$^{\square}$
ARC	0.25	Question (Q) + Candidate Answers (A)	0.435$^{\diamond}$ / 0.422$^{\square}$	Q	0.227$^{\diamond}$ / 0.227$^{\square}$
				A	0.245$^{\diamond}$ / **0.344**$^{\square}$
PIQA	0.5	Goal (G) + Solution 1 & 2 (Sol)	0.795$^{\diamond}$ / 0.794$^{\square}$	G	0.495$^{\diamond}$ / 0.495$^{\square}$
				Sol	**0.735**$^{\diamond}$ / **0.724**$^{\square}$
CSQA	0.2	Question (Q) + Candidate Answers (A)	0.738$^{\diamond}$ / 0.727$^{\square}$	Q	0.196$^{\diamond}$ / 0.196$^{\square}$
				A	**0.218**$^{\diamond}$ / 0.184$^{\square}$

Partial input training results (in accuracy). Scores above random are in bold. ◇: RoBERTa-Large; □: COMET(BART).

TextFooler

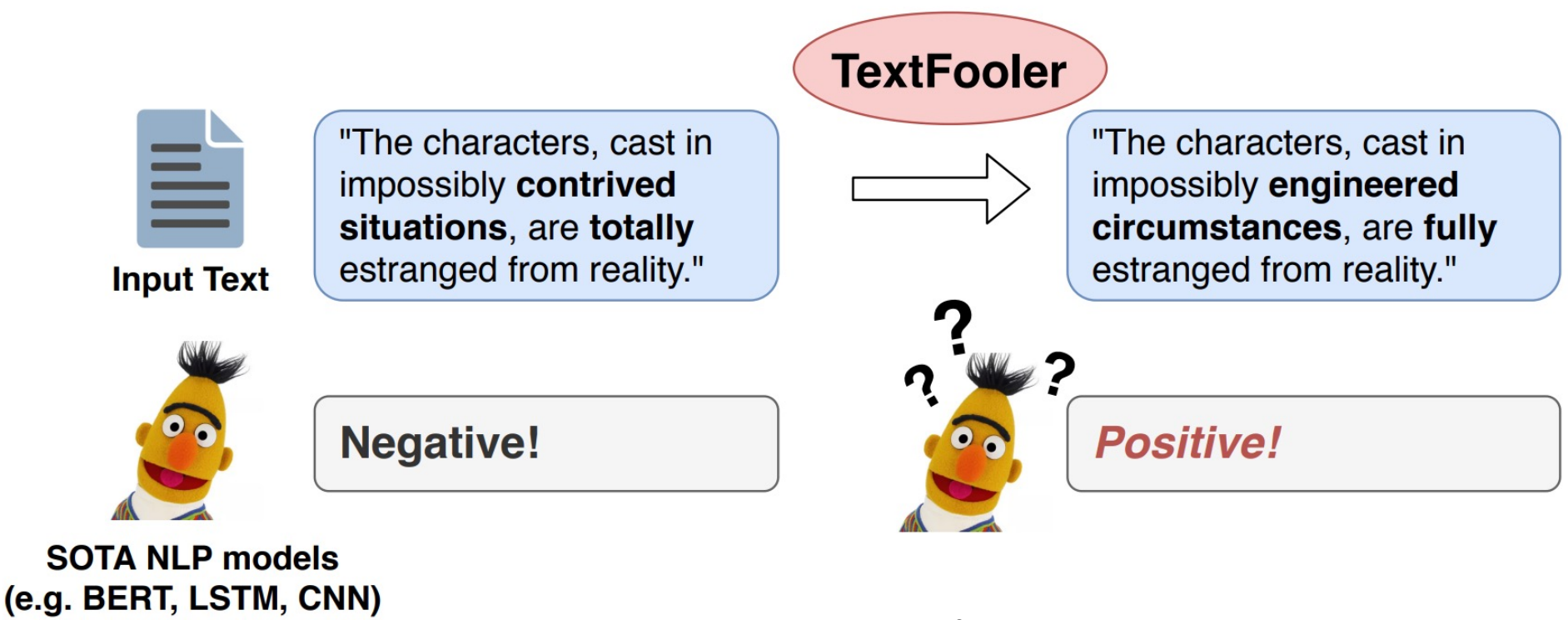


Figure adapted from Jin et al. 2020

Results & Discussion – Adversarial Attack

Question: Ira had to make up a lab investigation after school. He obtained the materials, chemicals, equipment, and protective gear from his teacher. Quickly, but cautiously, he conducted the steps in the written experiment procedure. To save time, he decided to record his observations and results later. Which will most likely be negatively affected by his decision?

Before

A: the ability to follow directions
B: the ability to write a valid report
C: the ability to follow the safety guidelines
D: the ability to come up with a conclusion

After

A: the capacity to follow directions
B: the ability to write a valid report
C: the ability to follow the safety guidelines
D: the ability to come up with a conclusion

Correct choice: B
Model's choice: B ✓
Model's choice after perturbation: A ✗

Results & Discussion – Adversarial Attack

- TextFooler (Jin et al., 2020) used as Adversarial Attack Algorithm.
- Both RoBERTa and COMET(BART) show brittleness.
- Sharper drop observed in ARCT, ARC and PIQA, the same tasks which are flagged in partial input training.

Task	Random	Model	Before	After	Δ	Δ%
ARCT	0.5	RoBERTa-Large	0.831	0.476	-0.355	42.7%
		COMET(BART)	0.795	0.512	-0.283	35.5%
ARC	0.25	RoBERTa-Large	0.435	0.157	-0.278	63.9%
		COMET(BART)	0.422	0.107	-0.315	74.7%
PIQA	0.5	RoBERTa-Large	0.795	0.306	-0.489	61.5%
		COMET(BART)	0.794	0.286	-0.508	64.0%
CSQA	0.2	RoBERTa-Large	0.738	0.536	-0.202	27.4%
		COMET(BART)	0.727	0.500	-0.226	31.3%

Results & Discussion – Data Contamination

- Based on methodology from GPT-3 Paper (Brown et al., 2020).
- ARCT is entirely clean.
- Different levels of contamination.
- When testing on a clean and dirty set, data contamination was found to have a negligible impact.
- Data contamination should not be a factor here.

Name	Split	Dirty Percentage	Clean Percentage
ARCT	test	0%	100%
ARC	test	1.19%	98.81%
PIQA	dev	13.22%	86.78%
CSQA	dev	5.08%	94.92%

Name	Split	Original Accuracy Score	Accuracy Score Dirty Set	Accuracy Score Clean Set
ARC	test	0.435	0.714 (+0.279)	0.432 (-0.003)
PIQA	dev	0.795	0.835 (+0.040)	0.789 (-0.006)
CSQA	dev	0.738	0.726 (-0.012)	0.739 (+0.001)

Name	Split	Original Accuracy Score	Accuracy Score Dirty Set	Accuracy Score Clean Set
ARC	test	0.422	0.643 (+0.221)	0.420 (+0.002)
PIQA	dev	0.794	0.819 (+0.025)	0.790 (-0.004)
CSQA	dev	0.727	0.710 (-0.017)	0.727 (+0.000)

Results & Discussion – Cross-Task Evaluation

- CSQA provides the best generalization. Also the least affected from the previous experiments.
- PIQA → ARC the only application below random baseline.
- ARCT weakest contributor.

	ARCT	ARC	PIQA	CSQA
ARCT	*0.831*	0.310	0.571	0.293
ARC	0.589	*0.435*	0.627	0.343
PIQA	0.597	**0.230**	*0.795*	0.552
CSQA	0.627	0.384	0.687	*0.738*
Random	0.5	0.25	0.5	0.2

Results & Discussion – Shortcut Exploration

- No major class imbalance that could explain the results.
- No strong **contiguous** lexical cues found.

Unigrams			Bigrams		
Unigram	Coverage (ξ_k)	Productivity (π_k)	Bigram	Coverage (ξ_k)	Productivity (π_k)
(to,)	**0.13**	**0.26**	(of, the)	0.07	0.15
(in,)	0.13	0.25	(in, the)	0.06	0.24
(of,)	0.13	0.25	(to, the)	0.04	0.24
(a,)	0.11	0.22	(amount, of)	0.03	0.25
(the,)	0.09	0.25	(from, the)	0.03	0.27
(water,)	0.09	0.15	(in, a)	0.03	0.30
(from,)	0.07	0.23	(on, the)	0.03	0.22
(and,)	**0.06**	**0.41**	(the, same)	0.02	0.30
(on,)	**0.06**	**0.26**	(number, of)	0.02	0.16
(for,)	**0.05**	**0.29**	(the, amount)	0.02	0.24
(an,)	**0.05**	**0.29**	(of, a)	0.02	0.25

Unigrams			Bigrams		
Unigram	Coverage (ξ_k)	Productivity (π_k)	Bigram	Coverage (ξ_k)	Productivity (π_k)
(a,)	**0.10**	**0.52**	(in, the)	0.03	0.41
(of,)	0.07	0.50	**(on, the)**	**0.03**	**0.54**
(to,)	0.07	0.49	**(of, the)**	**0.03**	**0.50**
(and,)	**0.07**	**0.52**	(with, a)	0.03	0.47
(in,)	0.06	0.47	**(use, a)**	**0.02**	**0.51**
(on,)	**0.06**	**0.53**	(to, the)	0.02	0.47
(the,)	0.05	0.40	**(in, a)**	**0.02**	**0.50**
(with,)	0.05	0.47	(and, then)	0.02	0.43
(it,)	0.05	0.48	**(into, the)**	**0.01**	**0.52**
(water,)	**0.04**	**0.52**	(top, of)	0.01	0.45
(your,)	0.04	0.45	(the, top)	0.01	0.47

Conclusion

- Models do better than expected without all input segments.
- Models appear to be **very** brittle when adversarially attacked.
- CSQA, the least affected task, generalizes better to other tasks.
- Data contamination tests show that memorization is not explaining factor.
- Tests performed to investigate two possible shortcuts did not find the explaining factor.

What now?

- Careful review dataset development methods and learning methods.
- Does this happen in challenges in general?
- Would be a problem if so!
- Less infatuation with metrics, more infatuation with comprehension?

THANK YOU, QUESTIONS?

rmbranco@fc.ul.pt

https://github.com/nlx-group/Shortcutted-Commonsense-Reasoning

https://rubenbranco.github.io/

U LISBOA | UNIVERSIDADE DE LISBOA

Ciências ULisboa Faculdade de Ciências da Universidade de Lisboa

Bibliography

- R. Geirhos, J.H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. ISSN 25225839.
- A. J. DeGrave, J. D. Janizek, and S.I. Lee. Ai for radiographic covid19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.
- S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.
- T. Niven and H.Y. Kao. Probing neural network comprehension of natural language arguments. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658-4664, Florence, Italy, July 2019. Association for Computational Linguistics.
- D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8018–8025, Apr. 2020.
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are fewshot learners. arXiv preprint arXiv:2005.14165, 2020.